Discovering Bioactive Peptides and Characterizing the Molecular Pathways that Control Their Activity

Professor Alan Saghatelian Andrew James Mitchell

Discovering Bioactive Peptides and Characterizing the Molecular Pathways that Control Their Activity

Abstract

Bioactive peptides constitute a major class of signaling molecules in animals and have been shown to play a role in diverse physiological processes, including hypertension, appetite and sleep. As a result, knowing the identity of these molecules and understanding the mechanisms by which they are regulated has basic and medical significance. In this dissertation, I describe the development and application of novel methods for discovering bioactive peptides and the molecular pathways that control their activity.

Recent analyses of mammalian RNAs have revealed the translation of numerous short open reading frames (sORFs). However, it is unknown whether these translation events produce stable polypeptide products that persist in the cell at functionally relevant concentrations. In Chapter 1, I describe a study in which we used a novel mass spectrometry-based strategy to directly detect sORF-encoded polypeptides (SEPs) in human cells. This analysis identified 115 novel SEPs, which is the largest number of mammalian SEPs discovered in a single study by more than a factor of 25. We observed widespread translation of SEPs from non-canonical RNA contexts, including polycistronic mRNAs and sORFs defined by non-AUG start codons. We also found that SEPs possess

properties characteristic of functional proteins, such as stable expression, high cellular copy numbers, post-translational modifications, sub-cellular localization, the ability to participate in specific protein-protein interactions and the ability to influence gene expression. Taken together, these findings provide the strongest evidence to date that coding sORFs constitute a significant human gene class.

In chapter 3, I describe a study in which we combine quantitative *in vivo* peptidomics, classical biochemical experiments and pharmacological studies in animal models to elucidate the metabolism of the neuropeptide substance P in the spinal cord. We identified two physiological substance P metabolites: the N-terminal fragments SP(1-9) and SP(1-7). Focusing our efforts on the SP(1-9)-producing pathway, we determined that an activity sensitive to the inhibitor GM6001 is the dominant SP(1-9)-generating activity in the spinal cord. We also show that GM6001 treatment causes a nearly three-fold increase in endogenous substance P levels in the spinal cords of mice, highlighting the functional relevance of the pathway blocked by this inhibitor.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I thank Alan Saghatelian for his insightful advice, both strategic and technical, on the research described herein; these projects would not have succeeded without his contributions. I also thank Alan for creating an excellent environment in which to work. The camaraderie I have felt with my colleagues has been a special part of my experience in graduate school and I don't think it would have been possible had Alan not actively fostered a culture of openness and collaboration in our group.

In addition, I thank Bogdan Budnik and John Neveau for providing technical advice on mass spectrometry experiments and good company at the Northwest happy hours; Arthur Tinoco for setting an excellent example in the laboratory and being a great collaborator; Whitney Nolte for answering so many of my silly questions; Anna Marie Lone for tolerating with good humor my somewhat frequent lectures on why America is better than Norway (and also answering my silly questions); Yui Vinayavekhin for ensuring that even the lab safety officer is not above the law; Edwin Homan for the breakroom conversations on politics and lipids; Tejia Zhang for inspiring me with her diligence in the laboratory and the neatness of her script; Yun-Gon Kim for his stoicism in the face of LTQ maintenance; Amanda McFedries for representing MCB; Adam Schwaid for his humor, competence and succinct emails; and, finally, Mathias Leidl and Jiao Ma for keeping the Saglab annex lively during the final stretch.

Chapter 1

Discovering and Characterizing Human Short Open Reading Frame-Encoded Polypeptides

Alan Saghatelian and I designed and developed the SEP-discovery platform with help from Drs. Amir Karger and James Cuff on generating the custom polypeptide databases and related support. Dr. Sarah Slavoff and I performed the peptidomics experiments. Dr. Bogdan Budnik analyzed the samples on the mass spectrometer. Dr. Sarah Slavoff, Adam Schwaid and I analyzed the data, synthesized peptides and cloned constructs. Adam Schwaid and I performed the quantitative SEP measurements and prepared samples for the L1000 experiment. Dr. Sarah Slavoff performed the heterologous expression experiments that defined the FRAT2-SEP start site and uncovered the FRAT2 protein-protein interaction with P32. Willis Read-Button and Aravind Subramanian performed the L1000 assay and subsequent gene set enrichment analysis. Joshua Levin generated the RNA-seq data. Moran Cabili and John Rinn assembled the K562 transcriptome from the RNA-seq data and culled the lincRNAs.

1.1 Introduction

Short open reading frames (sORFs) in the 5'-untranslated region (5'-UTR) of eukaryotic mRNAs (uORFs) are well studied (1) and several have been shown to produce polypeptides (2, 3). In addition to uORFs, a handful of other sORFs in bacteria (4), viruses (5), plants (6, 7), *Saccharomyces cerevisiae* (8), *Caenorhabditis elegans* (9), insects (10, 11), and humans (12) have been discovered to encode polypeptides. Notable among them are the peptides encoded by the polycistronic *tarsel-less* (*tal*) gene in *Drosophila*, which are as short as 11 amino acids and regulate fly morphogenesis (10, 11).

Recently, computational analysis of the mouse transcriptome using improved gene-prediction algorithms has suggested that sORF-encoded polypeptides (SEPs) are significantly underrepresented in current protein catalogues (13), leading to speculation that coding sORFs constitute an unrecognized mammalian gene class. This hypothesis was bolstered by subsequent ribosome profiling studies in mouse embryonic stem cells, which found evidence that hundreds of sORFs in the mouse transcriptome are engaged by the protein translation machinery. However, since these studies did not directly detect the presence of any sORF-encoded polypeptides (SEPs), it remains unclear whether sORFs are widely translated into polypeptides that persist in the cell at functionally relevant concentrations. Indeed, follow-up experiments on sORFs identified as being translated by ribosome profiling failed to identify any polypeptide expression (14), indicating that at least some of the implicated sORFs are false positives.

While no general approach exists for discovering SEPs, several attempts have been made to systematically identify these molecules. In *E. coli*, for example, experiments in which predicted sORFs were epitope-tagged revealed 18 SEPs (15). In another example, a combination of computational and experimental approaches identified 299 potentially coding sORFs in *S. cerevisiae*, four of which were confirmed to produce protein and 22 of which appeared to regulate growth (8). Finally, in human cells, an unbiased proteomics approach identified a total of four SEPs (defined here as polypeptides that are synthesized on the ribosome at a length of 150 amino acids or less) between the K562 and HEK293 cell lines with a length distribution of 88-148 amino acids (16).

The discordance between the small number of SEPs detected in human cells (16) and the large number of coding sORFs described by ribosome profiling (17) and computational methods (13) leaves open the possibility that SEPs are not produced as predicted by these methods or else are rapidly degraded and therefore not detectable. Indeed, one might speculate that sORFs are translated as a unique mode of gene regulation whereby the process of translation itself rather than the synthesized product is the functional agent, or as part of a "checkpoint" or stalling mechanism in the trafficking of functional RNAs, or perhaps simply due to stochastic binding of ribosomes to pseudo-initiation sites that arise in the genome by chance.

Thus, there is a need for an approach that can directly detect and validate the products of translated sORFs on a global scale. We therefore developed a novel

strategy for SEP discovery that integrates an optimized peptidomics and bioinformatics platform with a rigorous evaluation procedure based on manual spectra analysis and use of synthetic standards. Applying this approach to human cells, we uncovered 118 unannotated SEPs, which is the largest number of human SEPs ever reported in a single study by approximately a factor of thirty. We also analyze SEP-encoding sORFs to reveal several unexpected features of SEP translation, including widespread initiation at non-AUG start codons and polycistronic expression. Perhaps most intriguingly, though, we find that SEPs possess properties characteristic of functional proteins, such as stable expression, high cellular copy numbers, post-translational modifications, subcellular localization, the ability to participate in specific protein-protein interactions and the ability to influence gene expression. Taken together, these findings provide the strongest evidence to date that coding sORFs constitute a significant human gene class.

1.2 Peptidomics analysis of human myelogenous leukemia cells

Peptidomics is a liquid chromatography-tandem mass spectrometry (LC-MS/MS)-based technique aimed at the comprehensive visualization and analysis of endogenous peptides (18, 19). Peptidomics experiments are distinguished from traditional proteomics experiments in that the core workflow contains steps designed to preserve and enrich small polypeptides (18-21). In the context of the present study, the use of peptidomics is intended to increase the total number of

SEPs detected and expand the detection range to include shorter SEPs by eliminating the signal-suppressing effects of large proteins. We isolated peptides from K562 cells (22), a human leukemia cell line, because these cells were the subject of the most successful SEP discovery effort to date and we could therefore use the previously reported SEPs to benchmark our performance (16).

Endogenous peptides were isolated using an optimized peptidomics workflow developed in our laboratory (21) (Figure 1.1), with great care being taken to reduce proteolysis. Proteolysis is detrimental because the processing of cellular proteins greatly increases the complexity of the peptidome, which deteriorates the signal-to-noise ratio during the subsequent analysis (23). Practically, isolation of the peptidome from K562 cells is accomplished by boiling a frozen cell pellet to lyse the cells and simultaneously inactivate any proteases. Heat inactivation of proteolytic activity is common practice in peptidomics and has led to the identification of known peptides from cells or tissues to demonstrate its reliability (24-27). After cell lysis, the lysate is passed through a molecular weight cut-off filter to separate small polypeptides (hereafter, “peptidome”) from the rest of the proteome. We prepared two sets of samples, one enriched using a 30 kDa MWCO filter and another enriched using a 10 kDa MWCO filter. This was done to maximize our coverage of the smallest SEPs while ensuring that larger SEPs were not overlooked.

After isolating the peptidome, we reduced and alkylated the sample to eliminate disulfide bonds and prevent them from re-forming; covalent links

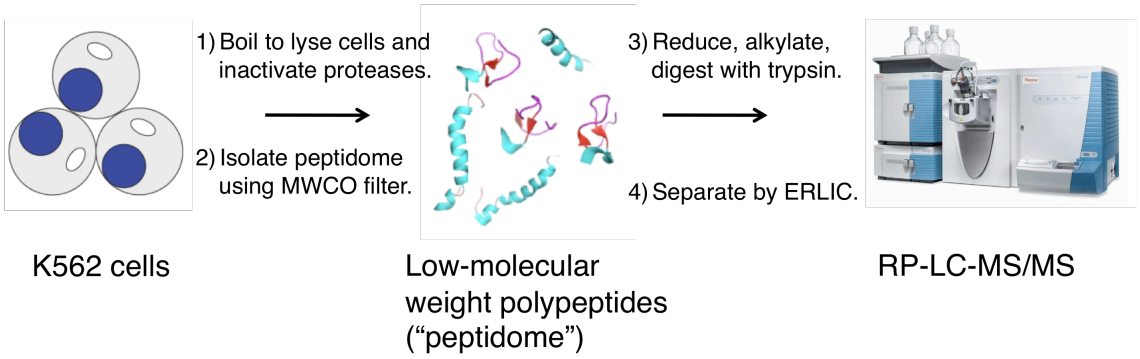


Figure 1.1 Peptidomics workflow. First, we boil a frozen K562 pellet to lyse the cells and simultaneously eliminate protease activity. Second, we isolate the peptidome by passing the cell lysate through a molecular weight-cutoff filter. Next, we reduce and alkylate the peptide mix to eliminate disulfide bonds and prevent them from re-forming prior to analysis; covalent links between peptide side chains can make definitive MS/MS spectra interpretation difficult or impossible. After that, we fractionate the peptidome by electrostatic repulsion-hydrophilic interaction chromatography (ERLIC), which separates analytes by hydrophobicity and isoelectric point. This step simplifies the sample to allow for deeper coverage of the peptidome. Finally, we analyze each ERLIC fraction by reversed-phase-liquid chromatography-tandem mass spectrometry (RP-LC-MS/MS), thus generating a collection of spectra from which the identities of the peptides can be determined.

between peptide side chains can make definitive MS/MS spectra interpretation difficult or impossible. The sample was then exposed to trypsin to generate peptide fragments that are ideal for subsequent LC-MS/MS analysis. However, prior to analysis, the trypsinized peptides were fractionated by electrostatic repulsion-hydrophilic interaction chromatography (ERLIC)(28), which separates peptides based on their hydrophobicity and isoelectric point. ERLIC fractionation has been reported to significantly improve detection sensitivity in proteomics experiments (29), so we included this step to deepen our peptidome coverage. Finally, each ERLIC fraction was analyzed by nano-flow reversed-phase (RP)-LC-MS/MS system with a high-resolution mass spectrometer.

1.3 Analyzing peptidomics data to identify novel SEPs

To identify SEPs, it was necessary to develop a modified protocol for LC-MS/MS data analysis. Standard proteomics and peptidomics approaches identify peptides by matching experimentally observed spectra to databases of predicted spectra based on annotated genes. Such databases would not necessarily contain the predicted spectra of SEPs. We therefore created a custom database containing all polypeptides that could possibly be translated from the annotated human transcriptome (The National Center for Biotechnology Reference Sequence, or RefSeq (30)) or the reverse-complement thereof (Figure 1.2A). We were interested in polypeptides that could be translated from the reverse-complement of annotated transcripts because of reports of pervasive transcription of the antisense strands of eukaryotic genes ((31)).

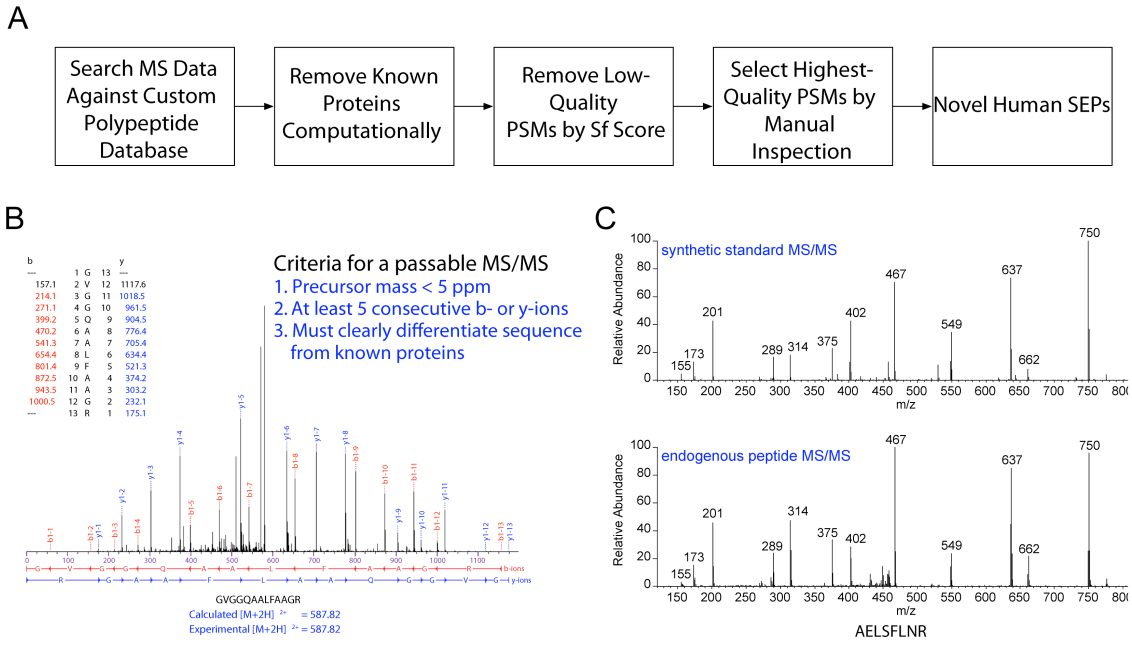


Figure 1.2 Discovering SEPs. (A) An LC-MS/MS-based peptidomics platform was used to profile K562 cells. The MS/MS data were searched against a custom protein database derived from human RefSeq transcripts to identify polypeptides in K562 cells. Tryptic peptides that were exact matches to a segment of an annotated protein were computationally filtered. In addition, tryptic peptides that differed from annotated proteins by only a single amino acid were also removed to avoid the false identifications arising from point mutations in known proteins. The sequence assignment of these putative SEPs was validated by visual inspection of the tandem MS spectra. Finally, we referenced K562 RNA-seq data to verify that that detected peptides were derived from a sORF rather than an unannotated ORF longer than 450 nucleotides or a mutated annotated ORF. Any tryptic peptide that fit these criteria was identified as arising from a novel human SEP. (B) Tandem MS spectra were visually inspected to ensure that there was sufficient sequence coverage to unambiguously differentiate the peptide from similar known protein sequences. The spectra were required to have a precursor mass error of less than 5 ppm and a sequence tag of five consecutive b- or y-ions. (C) We experimentally validated some of these assignments by chemically synthesizing the diagnostic peptide and comparing its tandem MS spectra of that of the endogenous peptide. This particular peptide is derived from a sORF found on a non-coding RNA (chr16:86563805-86589025).

Using Sequest, an analysis program used to identify peptides from MS/MS spectra (32, 33), we compared >200,000 MS/MS peptide spectra to this RefSeq-derived polypeptide database. This resulted in 6548 unique peptide identifications. We arrived at a tentative list of SEPs by keeping only those tryptic peptides that differed by at least two amino acids from every annotated protein to minimize the possibility of false positives arising from polymorphisms in annotated genes. We then winnowed the list down further to a candidate set by eliminating all peptide-spectrum matches (PSMs) with an Sf score of less than 0.4, which is a typical threshold used in proteomics studies to cull PSMs that can be used with confidence to identify proteins (34). (The Sf score is a composite metric generated by the SEQUEST algorithm to indicate the strength of a PSM. It takes into account the preliminary score, the cross-correlation and the difference in strength between the highest scoring PSM and the second highest scoring PSM for a given spectra.)

Due to the small size of SEPs, it is unlikely that an unbiased peptidomics experiment will detect more than one tryptic fragment of a given SEP. This contrasts with standard proteomics studies, which, on account of the numerous tryptic fragments generated from larger polypeptides, will typically uncover two or more peptides to support the presence of a protein. Realizing that we would likely not be able to rely on the confidence contributed by the inherent redundancy of multiple-peptide protein identifications for SEP discovery, we submitted the

candidate PSMs to a rigorous evaluation procedure to ensure the highest confidence in each peptide identification.

First, we discarded any PSM with an Sf score of less than 0.75. This eliminated over 95% of the candidate set. We then visually examined each remaining MS/MS spectrum to ensure that it met a stringent set of criteria (Figure 1.2B). In particular, we required that there be a sequence tag of five consecutive b- or y-ions, a precursor mass error of <5 ppm, and sufficient sequence coverage to unambiguously differentiate each peptide from every annotated protein sequence. This step reduced the remaining peptide pool by approximately 75%, for a total of 39 SEPs, three of which were previously reported and thus served as positive controls (16) and 36 of which are novel. Our PSM evaluation procedure therefore selected the highest quality ~1% of the peptide identifications in our original candidate set. As a check on the effectiveness of this procedure, we compared the experimentally collected MS/MS spectra of several identified peptides to that of identical synthetic peptides (Figure 1.2C). The spectra of the synthetic peptides were nearly identical to those of the endogenous peptides, confirming the identifications and validating our spectra evaluation procedure.

1.4 Probing the human transcriptome using deep RNA sequencing technology

Although we were extremely careful in evaluating peptide-spectra matches to ensure that the tryptic peptides we identified were present in the sample, it occurred to us that there were still several potential sources of false positives that we had not eliminated. For one thing, we were not able to exclude the possibility that some of the identified peptides were produced from a long ORF on an RNA that was not catalogued in RefSeq (30) rather than from a sORF on an annotated transcript, as we suspected was the case. Indeed, since it is probable that not all RefSeq genes will be expressed in a given cell line, we could not even be sure that the transcripts to which we had ascribed SEPs were present in K562. Additionally, a recent report had indicated that mRNAs can undergo post-transcriptional modifications that lead to alternate protein sequences (35). This phenomenon could also lead to the misidentification of SEPs. Lastly, although we had only admitted peptides whose sequence was at least two amino acids different from the nearest annotated protein sequence, it was still conceivable that a point mutation in an annotated gene could have led to an erroneous SEP identification.

Having a detailed and comprehensive knowledge of the K562 transcriptome would enable us to rule out these potential sources of false positives. We therefore deep-sequenced K562 cellular RNA and assembled it

into the complete K562 transcriptome (Figure 1.3). Poly-adenylated RNA and total RNA were isolated separately from cultured K562 cells and cDNA libraries were generated from each sample. These libraries were then sequenced using Illumina high-throughput sequencing technology and the resulting reads were aligned to the human genome using the splice junction mapper TopHat (36). Finally, Cufflinks (36) was used to assemble the transcriptome. Including mono-exonic RNAs, this analysis yielded over 700,000 unique transcripts.

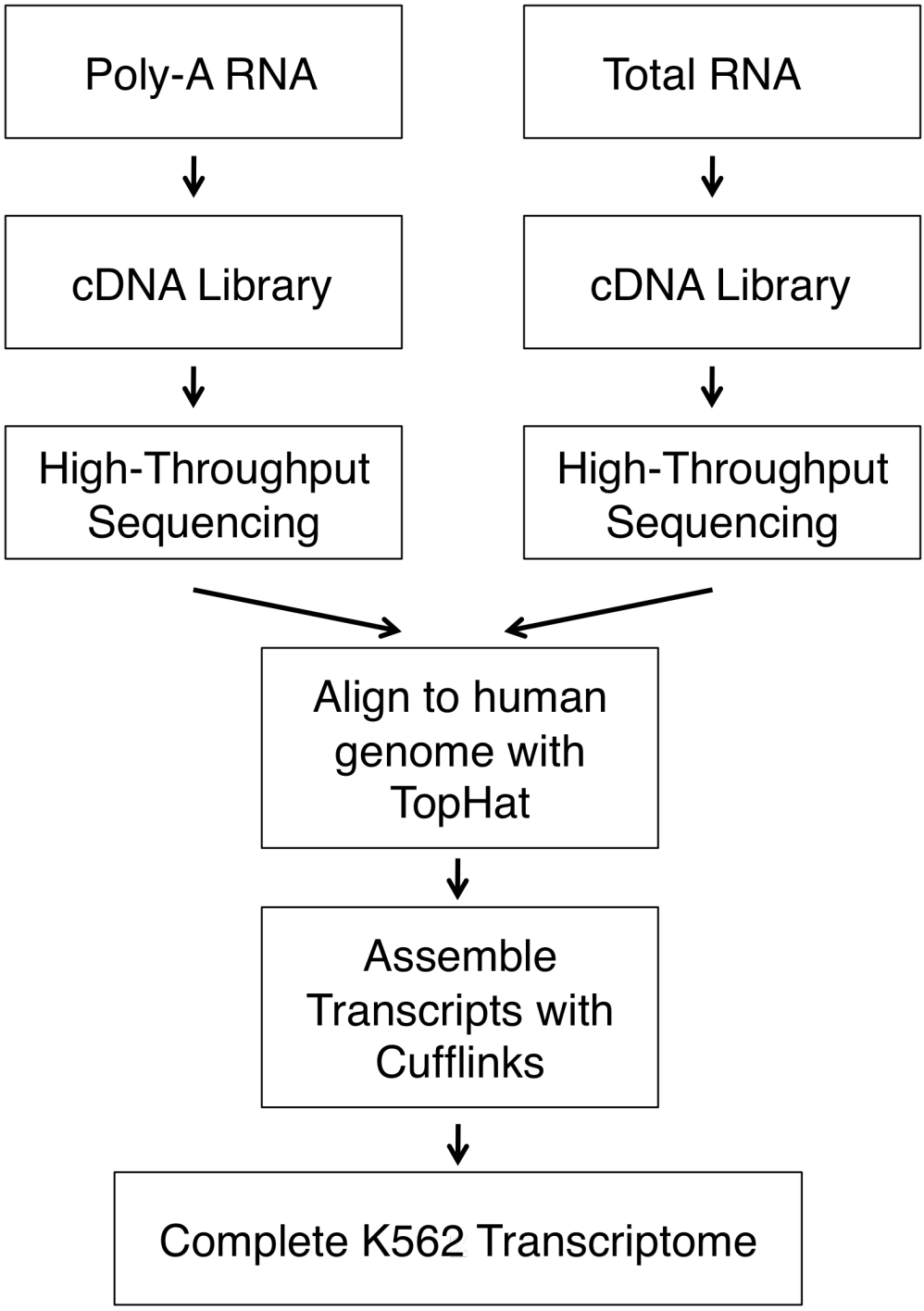


Figure 1.3 Generating the complete transcriptome of K562 cells. Poly-adenylated RNA and total RNA were isolated separately from cultured K562 cells and cDNA libraries were generated from each sample. These libraries were then sequenced using Illumina high-throughput sequencing technology and the resulting reads were aligned to the human genome using the splice junction mapper TopHat (36). Finally, Cufflinks (36) was used to assemble the transcriptome.

1.5 Validating SEPs encoded by annotated transcripts

We wanted to determine whether the 39 RefSeq transcripts to which we had assigned SEPs in our previous discovery effort were actually present in K562. Crosschecking these transcripts with our assembled RNA-seq data revealed that 37/39 SEPs were present. Interestingly, the implicated antisense transcript, the existence of which had never before been experimentally verified, was present in the sample.

Next, we wanted to verify that there were no long, unannotated ORFs that could be producing the detected SEPs. We therefore searched the 37 detected peptides that we had assigned to these transcripts against a theoretical protein database generated by *in silico* translation of the complete K562 transcriptome using the BLAST algorithm (37). All 37 SEPs mapped uniquely to the annotated transcript sORF to which they had been assigned in our initial study (Table 1.1). With this analysis, then, we eliminated the possibility that the detected peptides had arisen from point mutations in annotated genes, longer unannotated ORFs, or post-transcriptional modification or editing of RNAs and were thereby left with a high-confidence set of annotated transcript-derived SEPs.

The 37 SEPs discovered through analysis of RefSeq transcripts fall into five major categories: (i) those located in the 5'-UTR, (ii) those located in the 3'-UTR, (iii) those located (frameshifted) inside the main coding sequence (CDS), (iv) those located on non-coding RNAs (ncRNAs), and (v) those located on antisense transcripts (Figure 1.4). Many of these SEPs appear to be derived from

Table 1.1 List of high-confidence SEPs derived from RefSeq transcripts. When present, an upstream in-frame AUG was assumed to be the initiation codon. If no upstream AUG was present, the initiation codon was assigned to an in-frame near-cognate non-AUG codon embedded within a Kozak-consensus sequence (38). In a few cases, neither of these conditions was met, so the codon immediately following an upstream stop codon was used to determine maximal SEP length. Chromosome coordinates are from the University of California Santa Barbara Genome Browser, assembly H19.

Transcript Category	Detected Peptide	Sf Score	Start Codon	SEP Length (aa)	Chromosome Coordinates
RefSeq	AAPGALPEAAVGPR	0.81	ATG	96	chr9:139256352-139264369 strand=-
RefSeq	AGAPAVGLLLANER	0.93	GTG	39	chrX:16859470-16888534 strand=-
RefSeq	QLPPAAAVGDAGQLGR, APGGAAAGPGAPGCGGAGGQGPAPGGAAAAAAR	0.91, 0.98	ACG	103	chr10:99092201-99094454 strand=-
RefSeq	ATPGLQQHQQPPGPGR, ATPPGGTGHEGLSGGAADVASGVGSGR	0.92, 0.95	ATG	83	chr9:139557366-139565706 strand=+
RefSeq	IVVDELSNLK, QQQNSNIFFLADR, NILDELKK, EYQEIENLDK	0.9, 0.89, 0.74, 0.87	ATG	96	chr2:190526195-190535440 strand=+
RefSeq	TAPSSTATTASASCAATR	0.96	ATG	62	chr7:150646657-150675423 strand=-
RefSeq	LQVGPADTQPR	0.93	ATG	88	chr9:123612077-123639492 strand=-
RefSeq	STAACQTSSIATR	0.87	ATG	97	chr14:103800538-103809402 strand=+
RefSeq	GSSAAVGPR	0.84	stop	78	chr16:89574827-89607413 strand=+

Table 1.1 (Continued)

Transcript Category	Detected Peptide	Sf Score	Start Codon	SEP Length (aa)	Chromosome Coordinates
RefSeq	TAAAAAAGTITRPR	0.78	GTG	102	chr8:64080459-64125260 strand=+
RefSeq	GVGGQAALFAAGR, AGGDLPLQPQPGGAAAR, AAQAFFPAAELAQAGPER	0.96, 0.93, 0.96	GTG	88	chr8:144897399-144897840 strand=-
RefSeq	AVAAAAAAAPDPGGR	0.81	acg	91	chr10:98288128-98346562 strand=-
RefSeq	GGLGAASIAADGAPR	0.86	ctg	115	chr4:122737616-122745077 strand=-
RefSeq	SSTPAPPQGQFLPPSI	0.78	acg	74	chr7:100464771-100471014 strand=+
RefSeq	VAVEEGLPGDPVAER	0.94	acg	107	chr11:65686750-65689023 strand=+
RefSeq	EGSVHPQVE	0.76	atg	87	chr10:101992055-102005758 strand=-
RefSeq	GAIGGGGAGVQGQTAGAR	0.91	atg	143	chr5:180650039-180662529 strand=+
RefSeq	VAAVAVGSQAVLQILSR	0.9	atg	77	chr19:42713286-42721897 strand=-
RefSeq	WTSSTSSPNTSGAPR	0.94	atg	77	chr19:12949331-12969791 strand=+
RefSeq	NPPLVQDTVSGK	0.9	atg	111	chr1:150522391-150532570 strand=+
RefSeq	QTAFGKWYESLLNNR	0.78	stop	63	chr3:193363602-193386115 strand=+

Table 1.1 (Continued)

Transcript Category	Detected Peptide	Sf Score	Start Codon	SEP Length (aa)	Chromosome Coordinates
RefSeq	TAAAAAAGTITRPR	0.78	GTG	102	chr8:64080459-64125260 strand=+
RefSeq	GVGGQAALFAAGR, AGGDLPLQPQPGGAAAR, AAQAFFPAAELAQAGPER	0.96, 0.93, 0.96	GTG	88	chr8:144897399-144897840 strand=-
RefSeq	AVAAAAAAAPDPGGR	0.81	acg	91	chr10:98288128-98346562 strand=-
RefSeq	GGLGAASIAADGAPR	0.86	ctg	115	chr4:122737616-122745077 strand=-
RefSeq	SSTPAPPQGQFLPPSI	0.78	acg	74	chr7:100464771-100471014 strand=+
RefSeq	VAVEEGLPGDPVAER	0.94	acg	107	chr11:65686750-65689023 strand=+
RefSeq	EGSVHPQVE	0.76	atg	87	chr10:101992055-102005758 strand=-
RefSeq	GAIGGGGAGVQGQTAGAR	0.91	atg	143	chr5:180650039-180662529 strand=+
RefSeq	VAAVAVGSQAVLQILSR	0.9	atg	77	chr19:42713286-42721897 strand=-
RefSeq	WTSSTSSPNTSGAPR	0.94	atg	77	chr19:12949331-12969791 strand=+
RefSeq	NPPLVQDTVSGK	0.9	atg	111	chr1:150522391-150532570 strand=+
RefSeq	QTAFGKWYESLLNNR	0.78	stop	63	chr3:193363602-193386115 strand=+
RefSeq	TWLPSCEDLTLPGGR	0.92	atg	50	chr7:158799724-158814542 strand=+

Table 1.1 (Continued)

Transcript Category	Detected Peptide	Sf Score	Start Codon	SEP Length (aa)	Chromosome Coordinates
RefSeq	AVAGAAAGAGGR	0.79	atg	73	chr19:13059508-13067950 strand=-
RefSeq	AEEQPGLGPGAAGR	0.94	atg	149	chr7:100032962-100034242 strand=-
RefSeq	RAVPAQGLLQSTPTCMPWTP	0.84	atg	54	chr1:160061156-160064154 strand=-
RefSeq	NTTQESLEKGP	0.78	stop	32	chr22:41740383-41756157 strand=+
RefSeq	EALNEFLTR	0.87	stop	22	chr4:169908762-169911558 strand=-
RefSeq	AEPLQTAGQAGR	0.83	atg	59	chr11:118964597-118966163 strand=-
RefSeq	AGNLILLQ	0.82	stop	23	chr3:124945640-125042272 strand=-
RefSeq	STTIGGMNQR	0.77	atg	26	chr12:48732236-48745011 strand=-
RefSeq	ERPANSLIDQCSQR	0.8	atg	54	chr2:131130309-131132956 strand=+
RefSeq	VFFKNLLAFAR	0.9	stop	22	chr6:80194734-80199064 strand=-
RefSeq	AELSFLNR	0.84	atg	70	chr16:86563805-86589025 strand=-
RefSeq	LLPLGASPAGVVGGGLAPPR	0.93	atg	85	chr22:21368073-21368526 strand=-
RefSeq	SLSSYGACSR	0.89	stop	71	chr17:35441928-35444379 strand=+

Table 1.1 (Continued)

Transcript Category	Detected Peptide	Sf Score	Start Codon	SEP Length (aa)	Chromosome Coordinates
RefSeq	FLPVDLSLLR	0.78	atg	90	chr1:155532795-155708399 strand=+
RefSeq	GPSGTQEMGPLSR	0.95	atg	102	chr19:3610043-3626771 strand=-

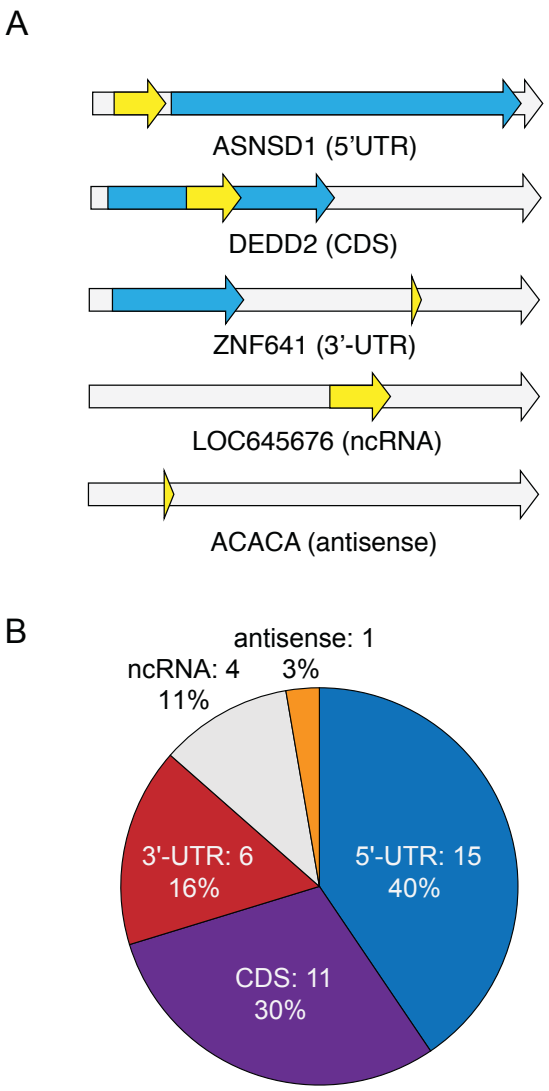


Figure 1.4 Overview of SEPs. (A) RNA maps illustrating the major categories of sORFs that are translated into SEPs. Categories include sORFs in the 5' untranslated region (5'UTR), sORFs in the coding sequence (CDS), sORFs in the 3' untranslated region, sORFs on non-coding RNAs (ncRNAs), and sORFs on transcripts antisense to annotated RefSeq transcripts.The gray arrow represents the RNA, the blue arrow represents annotated protein CDS (if present), and the yellow arrow represents the sORF. The numbers mark the boundaries of each element of the transcript in nucleotide bases from the 5' end of the transcript.
(B) Incidence of SEPs by sORF category.

polycistronic mRNAs, which is interesting because this phenomenon has historically been thought to be rare in eukaryotes. However, our findings here are consistent with those of ribosome profiling studies (17).

1.6 Discovering SEPs encoded by long intergenic non-coding RNAs (lincRNAs)

One intriguing feature of our analysis of RefSeq RNAs was the discovery of coding regions within transcripts that are annotated as non-coding. Since 2002, when long non-coding RNAs were first established as a transcriptional class through the sequencing of full-length cDNA libraries in mouse (39), interest in the functional properties of these molecules has grown rapidly. Commonly defined as transcripts that may posses mRNA-like properties but which lack ORFs longer than 100 amio acids, several dozen lncRNAs have now been shown to act as regulators of diverse cellular processes in mammals. For example, *XIST* contributes to X chromosome inactivation in by coating the inactivated chromosome (40); *HOTAIR* and *COLDAIR* interact with multiple protein complexes, including the polycomb-group protein Polycomb Repressive Factor 2 (PRC2), to control skin development (41, 42); and *H19* and *KCNQ1OT1* interact with chromatin and a host of protein complexes to effect imprinting (43, 44). However, although RNAs such as these are known to play a direct, mechanistic role in carrying out their functions, there are numerous other putatively functional lncRNAs whose specific role has not been elucidated. In these cases, it is not

clear whether the transcript possesses an intrinsic function or rather performs through its coding potential[1].

The question of whether all transcripts that meet the traditional lincRNA definition are in fact non-coding has spawned efforts to discern *a priori* between novel SEP-encoding transcripts and bona fide non-coding RNAs. Notably, Lin et al have developed a comparative genomics-based gene prediction algorithm called PhyloCSF, which outperforms other available algorithms at identifying small protein-coding regions in genomic sequences (49). However, recognizing that some coding sORFs will be too small for even sensitive algorithms to identify and also that many coding sORFs may be recently evolved, we wondered whether our experimental approach to SEP discovery might make a nice complement to computational approaches.

Recognizing that a majority of putative lincRNAs are not listed in RefSeq, we generated an extensive catalogue of nominally non-coding RNAs from our K562 RNA-seq data according to a previousy published protocol (50) (Figure 1.3). Our lincRNA culling protocol was as follows. First, we removed any transcripts with a non-lincRNA annotation. Second, we scanned each transcript in all three frames to identify regions that could code for one of the protein domains catalogued in the protein family database Pfam (51). Transcripts

[1] It is important to note, however, that these possibilities are not exclusive. In *E. Coli*, for example, the sugar transport-related sRNA (SgrS) transcript helps cells recover from glucose-phosphate stress by base-pairing with the mRNA *ptsG* to bring about its degradation via RNase E (45, 46) and also by encoding a 43 amino acid functional peptide called SgrT that prevents glucose uptake, probably by regulating the glucose transporter, PtsG (47, 48)

containing such regions were removed. We then removed all transcripts containing an ORF that appeared to be conserved at the protein coding level, as indicated by a positive phylogenetic codon substitution frequency score (PhyloCSF) (49). From the remaining pool of transcripts, we selected those that were multiexonic and originated from intergenic regions of the genome; these composed our lincRNA pool. We then generated a theoretical protein database by *in silico*-translating these transcripts in three frames and used this database to analyze our peptidomics datasets.

Ribosome profiling experiments in mouse cells indicate the presence of translated sORFs on nearly half of the lincRNAs analyzed (17), which is much higher than expected (52-54). By contrast, our peptidomics analysis identified 10 SEP-encoding lincRNAs (Table 1.2), which represents just 0.5% of the1866 lincRNAs detected in our RNA-seq analysis of K562. This disparity may result from a number of factors, including false positive identifications by ribosome profiling techniques of (14). Additionally, ribosome profiling may identify rare translational events that do not generate enough protein to be detected by LC-MS/MS, since mass spectrometry is biased towards the detection of more abundant peptides (55). It is also possible that some of the sORFs identified by ribosome profiling may produce polypeptides that are rapidly degraded and therefore would be undetectable using any analytical approach. Finally, the disparity may be a consequence of the fact that the study in question significantly undersampled lincRNAs: only 30 lincRNAs were examined when mammalian

Table 1.2 List of high-confidence SEPs derived from lincRNA transcripts. When present, an upstream in-frame AUG was assumed to be the initiation codon. If no upstream AUG was present, the initiation codon was assigned to an in-frame near-cognate non-AUG codon embedded within a Kozak-consensus sequence (38). In a few cases, neither of these conditions was met, so the codon immediately following an upstream stop codon was used to determine maximal SEP length. Chromosome coordinates are from the University of California Santa Barbara Genome Browser, assembly H19.

Transcript Category	Detected Peptide	Sf Score	Start Codon	SEP Length (aa)	Chromosome Coordinates
lincRNA	THLGTEGQCDLPGAGGPAR	0.98	stop	100	chr10:11925853-11937442 strand=
lincRNA	CPFVLLMSSMILLR	0.81	STOP	33	chr10:11980633 2-119859641 strand=+
lincRNA	QVLITNKNQ	0.81	ATG	29	chr11:65266565-65274602 strand=-
lincRNA	QRIPCVVILTK	0.84	stop	73	chr19:23278060-23286908 strand=+
lincRNA	KTLPMMGMIR	0.75	stop	30	chr2:107137814-107160732 strand=+
lincRNA	QVNEETLK	0.78	stop	143	chr3:107852804-107857456 strand=+
lincRNA	KNLFQNTSR	0.79	stop	59	chr4:10069715-10074643 strand=-
lincRNA	LDMNPKK	0.79	ATG	60	chr5:67726254-67730308 strand=-
lincRNA	IYQEEKK	0.84	stop	75	chr7:26438339-26538594 strand=+
lincRNA	RAGYSELE	0.87	ATG	69	chr7:96251318-96293650 strand=-

cells typically contain thousands of such transcripts. Future work coupling ribosome profiling with mass spectrometry should help resolve these questions and provide a better understanding of the factors governing SEP expression.

1.7 Discovering SEPs encoded by unannotated transcripts

Recognizing that the RefSeq- and lincRNA-derived databases did not contain every SEP that could be encoded by transcripts present in K562, we generated a theoretical protein database by *in silico*-translating all of our K562 RNA-seq transcripts. Searching this database against our data sets and subjecting the resulting PSMs to our evaluation workflow yielded an additional 66 novel, high-confidence SEPs (Table 1.3).

Table 1.3 List of high-confidence SEPs derived from K562 RNA-seq transcripts. When present, an upstream in-frame AUG was assumed to be the initiation codon. If no upstream AUG was present, the initiation codon was assigned to an in-frame near-cognate non-AUG codon embedded within a Kozak-consensus sequence (38). In a few cases, neither of these conditions was met, so the codon immediately following an upstream stop codon was used to determine maximal SEP length. Chromosome coordinates are from the University of California Santa Barbara Genome Browser, assembly H19.

Transcript Category	Detected Peptide	Sf Score	Start Codon	SEP Length (aa)	Chromosome Coordinates
non-annotated	APEPGAVLAPAEVVLR	0.95	agg	119	chr22:47048295-47073068 strand=+
non-annotated	NALQQENHILDGVK	0.96	stop	52	chr15:91565384-91574477 strand=+
non-annotated	LLVSGSPSAETLPLR	0.94	atg	128	chr5:34914296-34925392 strand=+
non-annotated	ALAQGSLTPSQIYSA	0.91	aag	52	chr22:17092426-17095991 strand=+
non-annotated	LSAPQPGPDILQAPAR	0.81	GTG	89	chr19:54693858-54697432 strand=+
non-annotated	VYIFQPVFEQYAK	0.92	atg	54	chr15:55609385-55613829 strand=-
non-annotated	NEQTELLYNK	0.9	stop	18	chr12:118649944-118650075 strand=+
non-annotated	ILEDFLPPSSSRPQS	0.84	stop	42	chr2:85132483-85133801 strand=+
non-annotated	DLPGVAPPRPSLSLSGP	0.83	atg	65	chr9:130209955-130216851 strand=-
non-annotated	AAASGQPRPEMQCPAEQTEIK	0.81	atg	58	chr5:14664778-14699800 strand=+

Table 1.3 (Continued)

Transcript Category	Detected Peptide	Sf Score	Start Codon	SEP Length (aa)	Chromosome Coordinates
non-annotated	KINIEIR	0.8	ctg	46	chr5:179660625-179718930 strand=-
non-annotated	KLEITSI	0.88	stop	25	chr5:108191309-108191755 strand=+
non-annotated	KLQLQC	0.75	stop	120	chr20:55981949-55984389 strand=-
non-annotated	KLSLLEL	0.79	stop	47	chr5:33479130-33479598 strand=+
non-annotated	KLVSEIK	0.78	stop	17	chr20:51270125-51270250 strand=-
non-annotated	KNILEPK	0.89	stop	15	chr16:11961991-11972092 strand=+
non-annotated	KPLEPLL	0.78	agg	26	chr14:55178788-55179023 strand=+
non-annotated	KQGGFVQVSANAL	0.75	atg	136	chr22:32014633-32026837 strand=-
non-annotated	KYPPPPP	0.81	stop	20	chr14:75201541-75205240 strand=-
non-annotated	LNINQSIAVSTATQR	0.96	AGG	55	chr2:200322928-200323580 strand=+
non-annotated	LPGQATTQQTFDQR	0.88	stop	54	chr19:56165091-56185542 strand=+
non-annotated	LVSAVLAGKE	0.75	CTG	43	chr1:7863564-7864928 strand=-
non-annotated	MNFILK	0.75	atg	49	chr3:44912513-44913077 strand=+

Table 1.3 (Continued)

Transcript Category	Detected Peptide	Sf Score	Start Codon	SEP Length (aa)	Chromosome Coordinates
non-annotated	PAVAAATLHLPAAPEGPH	0.78	atg	49	chr7:100169852-100183655 strand=-
non-annotated	QELIGASLHTAR	0.75	stop	119	chr1:228544743-228549628 strand=+
non-annotated	RIQVEQTR	0.81	atg	63	chr9:15055076-15056573 strand=+
non-annotated	RSVFPLLK	0.8	stop	22	chr10:69681657-69833652 strand=-
non-annotated	TSDAPRPSATPPGADPLNSAGPGAR	0.81	stop	103	chr19:55737961-55770381 strand=-
non-annotated	VTSWDGQNPPR	0.76	ATG	50	chr12:12966292-12982891 strand=+
non-annotated	AAPGPTAAAAAQASAAAR	0.82	CTG	108	chr2:231577583-231685792 strand=+
non-annotated	RLLIPPEK	0.82	stop	45	chrX:5214450-5216144 strand=+
non-annotated	SPTTDSYGIPQGCK	0.89	stop	40	chr1:175913973-176153786 strand=-
non-annotated	APLLVKD	0.75	stop	15	chr10:97772573-97772956 strand=+
non-annotated	PDEIIFK	0.75	stop	50	chr2:176785921-176794931 strand=+
non-annotated	DYILSLEMFSILLWG	0.77	stop	33	chr3:88101102-88108113 strand=-

Table 1.3 (Continued)

Transcript Category	Detected Peptide	Sf Score	Start Codon	SEP Length (aa)	Chromosome Coordinates
non-annotated	EDNFILK	0.77	CTG	37	chr7:155093676-155102099 strand=-
non-annotated	LNLYEIK	0.78	stop	52	chr12:104344401-104350979 strand=+
non-annotated	KIIYDK	0.78	ATG	35	chr3:173924415-173924516 strand=+
non-annotated	FGGFSLK	0.79	stop	63	chr15:48995625-48997517 strand=+
non-annotated	HGHSFPDPGLLLQNQGD	0.79	stop	122	chr7:66386236-66423532 strand=+
non-annotated	FEIFGEK	0.8	stop	37	chr4:164444822-164451827 strand=+
non-annotated	HDASSSPLGPPR	0.8	stop	55	chr16:87435666-87438903 strand=-
non-annotated	EEAYFR	0.83	stop	23	chr1:95657105-95663161 strand=-
non-annotated	CLVYVLDLITDACTIKPLFNK	0.86	stop	43	chr9:130128866-130129660 strand=+
non-annotated	ASPGEAGPAGGAAAGQGAPR	0.89	stop	73	chr1:16905808-16970994 strand=-
non-annotated	GAWGGGQLATAGSGPGQR	0.96	ATG	70	chr17:62205639-62207524 strand=-
non-annotated	DTEVLINTMSK	0.79	ATT	27	chr1:4036227-4073316 strand=+
non-annotated	VYKWLLCNVE	0.78	ATG	41	chr1:157243513-157253900 strand=+

Table 1.3 (Continued)

Transcript Category	Detected Peptide	Sf Score	Start Codon	SEP Length (aa)	Chromosome Coordinates
non-annotated	KPVFLLLLSIR	0.85	STOP	32	chr11:3532972-3542051 strand=+
non-annotated	FIPTEAWYSAGR	0.79	ATG	86	chr11:82783129-82805398 strand=+
non-annotated	IKFLLAPEENK	0.86	ATG	43	chr16:3054772-3058645 strand=+
non-annotated	FYPDYIK	0.77	TTG	22	chr11:12970531-13011090 strand=-
non-annotated	QMSSNILK	0.76	stop	50	chr15:31008518-31061502 strand=+
non-annotated	VAHENYMKFK	0.82	stop	59	chr21:35345400-35353552 strand=+
non-annotated	GIALGDIPNAR	0.94	GTG	18	chr6:68590370-68642035 strand=+
non-annotated	VLLDQHQR	0.8	stop	23	chr6:141167131-141219546 strand=-
non-annotated	YYELQRGTR	0.84	AAG	43	chr15:59060273-59063173 strand=-
non-annotated	GEMERGEIK	0.81	ATG	18	chr17:41373439-41383338 strand=-
non-annotated	CQDILEAGKR	0.85	ATC	70	chr19:23441500-23457032 strand=-
non-annotated	DLGSPMLK	0.76	ATG	52	chr2:23598100-23604170 strand=-
non-annotated	TASPYSRPE	0.75	ATG	58	chr2:66653867-66660602 strand=-

Table 1.3 (Continued)

Transcript Category	Detected Peptide	Sf Score	Start Codon	SEP Length (aa)	Chromosome Coordinates
non-annotated	LTVAGQGR	0.75	ATG	66	chr20:4173737-4176599 strand=+
non-annotated	SPFWAGQGQSR	0.85	GTG	104	chrX:118425492-118469573 strand=+
non-annotated	NLAGGSGLIP	0.76	stop	41	chrX:1515320-1517852 strand=-
non-annotated	AAALQFDLR	0.94	stop	23	chr21-35303432-35308177 strand=+
non-annotated	AQHGVHSNTASPGLPAGAPR	0.96	agg	66	chr7:150778180-150780257 strand=-

1.8 Identifying post-translationally modified SEPs

Functional proteins are commonly phosphorylated as a means of controlling their activity and many bioactive peptides are amidated at the c-terminus to prevent degradation by exopeptidases. We wondered whether SEPs undergo the same post-translational modifications. To answer this question, we reanalyzed our peptidomics data sets against the custom databases we generated from human RefSeq transcripts and K562 RNA-seq data using the SEQUEST (32) algorithm with specialized parameters designed to enable the identification of phosphorylated or c-terminally amidated peptides. While we did not detect post-translational modification of any of the SEPs we had previously discovered, the analysis yielded five additional novel SEPs that are post-translationally modified (Table 1.4). All of these SEPs were phosphorylated and one was also c-terminally amidated.

Table 1.4 List of high-confidence post-translationally modified SEPs derived from K562 RNA-seq transcripts. When present, an upstream in-frame AUG was assumed to be the initiation codon. If no upstream AUG was present, the initiation codon was assigned to an in-frame near-cognate non-AUG codon embedded within a Kozak-consensus sequence (38). In a few cases, neither of these conditions was met, so the codon immediately following an upstream stop codon was used to determine maximal SEP length. "@" indicates that the preceding residue is phosphorylated and "[" indicates that the terminus is amidated. Chromosome coordinates are from the University of California Santa Barbara Genome Browser, assembly H19.

Transcript Category	Detected Peptide	Sf Score	Start Codon	SEP Length (aa)	Chromosome Coordinates
non-annotated	VTLNLFLTS@IK	0.85	stop	49	chr21:38122898-38126719 strand=+
non-annotated	SLGGILFTIIS@K	0.93	CTG	30	chr16:89735690-89738512 strand=-
non-annotated	IFLITIQDFIIAVIIVHS@T@DSLQRLV	0.79	CTG	147	chr15:77471130-77474523 strand=+
non-annotated	MLNFILIS@ILERA	0.75	ATG	14	chr9:139557366-139565706 strand=+
non-annotated	PGIGAGTPVGPKVVGS@L[	0.80	stop	35	chr12:3949920-3950581 strand=-

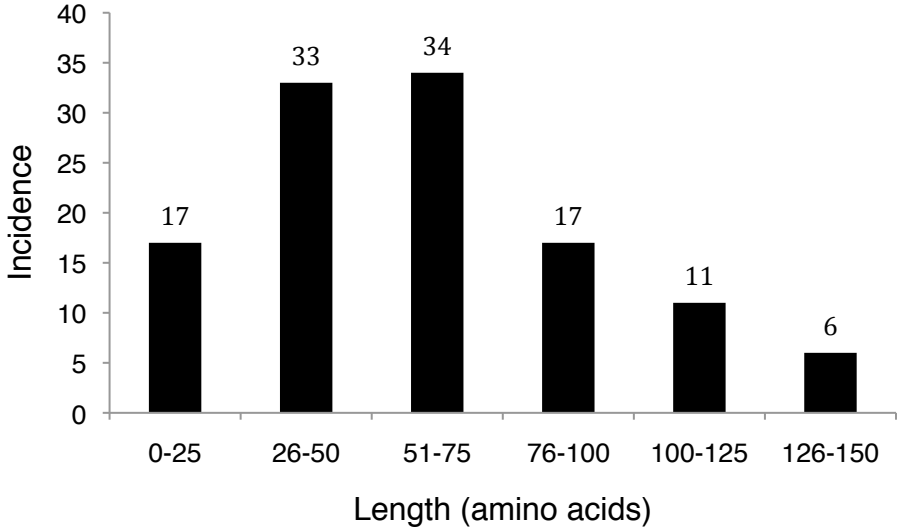
1.9 Exploring the global properties of SEPs

In total, we discovered 118 unannotated SEPs, three of which were previously reported and thus served as positive controls (16), and 115 of which are novel. This is the largest number of SEPs ever reported in a single study by approximately a factor of 30, which demonstrates the superior coverage afforded by our approach.

Wishing to explore the global properties of SEPs, we examined the size distribution and start codon usage of the molecules we had discovered. Because we perform our peptidomics analysis on trypsin-digested samples, we do not obtain full protein-level SEP sequence coverage and in particular do not directly observe the N terminus. We therefore used the following convention when assigning start sites to SEP-encoding sORFs. When present, the upstream-most in-frame AUG was assumed to be the initiation codon. If no upstream AUG was present, the initiation codon was assigned to an in-frame near-cognate non-AUG codon embedded within a Kozak-consensus sequence (38). In a few cases, neither of these conditions was met, so the codon immediately following an upstream stop codon was used to determine maximum SEP length.

Using this approach, we estimated the SEPs to range in length between 14-149 amino acids, with the majority (>70%) being <75 amino acids (Figure 1.5A). If we take a more conservative approach by using an AUG-to-stop or upstream-stop-to-stop, we obtain a similar SEP length distribution and retain our smallest SEPs, including the 14-mer (Figure 1.5B). As the shortest human SEP

A



B

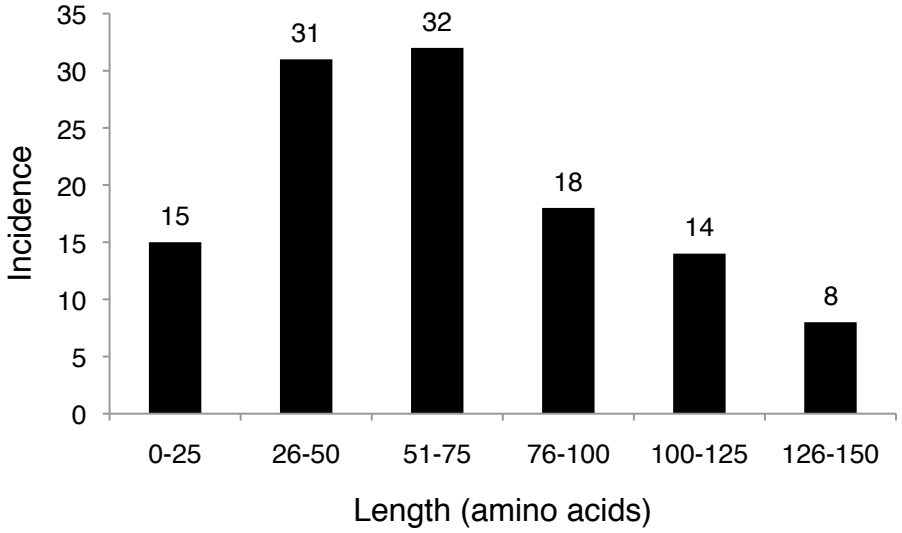


Figure 1.5 Length distribution of high-confidence SEPs. (A) SEP length distribution estimated by defining sORFs as follows: when present, an upstream in-frame AUG was assumed to be the initiation codon. If no upstream AUG was present, the initiation codon was assigned to an in-frame near-cognate non-AUG codon embedded within a Kozak-consensus sequence (38). In a few cases, neither of these conditions was met, so the codon immediately following an upstream stop codon was used to determine maximal SEP length. (B) Length distribution determined by defining sORF initiation sites as the codon immediately 3' of the stop codon upstream of the detected peptide unless an AUG was present, in which case the upstream-most AUG was defined as the start.

previously identified by mass spectrometry was 88 amino acids long (16), it is clear that our approach provides superior coverage of small SEPs. This is significant because many previously characterized, functional SEPs are under 50 amino acids (4, 10-12).

Another interesting feature of our results is the preponderance of non-canonical translation start sites: 62% of the detected SEPs do not initiate at AUG codons (Figure 1.6). This finding is consistent with the results of ribosome profiling experiments in mouse, which indicate that, globally, most ORFs contain non-AUG start sites (17). In addition, the human SEP start codon usage distribution we propose is similar to that observed by ribosome profiling in mouse (e.g. CUG is the second most common used codon in both data sets) (17).

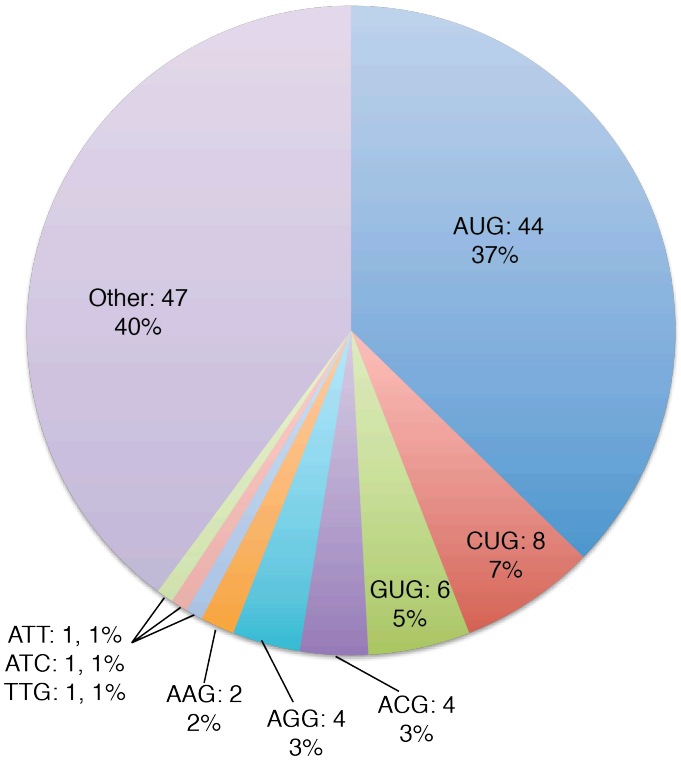


Figure 1.6 Probable start codon-usage distribution of SEP-encoding sORFs. Codon usage was estimated using the following sORF-defining convention: When present, an upstream in-frame AUG was assumed to be the initiation codon. If no upstream AUG was present, the initiation codon was assigned to an in-frame near-cognate non-AUG codon embedded within a Kozak-consensus sequence (30).

1.10 Testing SEP expression from RefSeq transcripts

Though in vitro experiments have shown that the eukaryotic translation machinery is capable of initiating translation at non-AUG codons and ribosome profiling experiments have indicated that non-AUG initiation is common in mouse, the preponderance of non-canonical translation start sites among the discovered SEPs is striking (Figure 1.6). We therefore wished to verify that our SEPs could be translated from non-AUG start sites. Moreover, because a majority of the discovered SEPs appear to arise from polycistronic transcripts, which is a phenomenon thought to be rare in eukaryotes, we sought to verify that the implicated full-length annotated transcripts of SEPs with AUG starts were competent to produce SEPs.

Constructs were designed to produce full-length mRNAs, including 5' and 3' UTRs, that matched those in the RefSeq database (56). We selected sORFs in the 5'-UTR, the 3'-UTR, or frameshifted within the CDS, and encoded a FLAG epitope tag at the 3'-end of each sORF (so that initiation is unperturbed). The uORFs *ASNSD1-SEP*, *PHF19-SEP, FRAT2-SEP*, *YTHDF3-SEP* and *EIF5-SEP* all produced cytoplasmically localized polypeptides, as detected by anti-FLAG immunofluorescence in transfected HEK293T cells (Figure 1.7A). (We refer to SEPs by appending "-SEP" to the name of the annotated CDS nearest the sORF; the sORF is given the same name but italicized.) Importantly, the fact that *FRAT2-SEP* and *YTHDF3-SEP*, which do not have upstream AUG codons,

Figure 1.7 Expression of SEPs. (A) Transient transfection of HEK293T cells with constructs containing a cDNA sequence corresponding to the full-length RefSeq mRNA (i.e., including the 5'- and 3'-UTRs). We appended a C-terminal FLAG-tag on the SEP coding sequence that could be detected by immunofluorescence. In these images the nuclei are stained with DAPI (blue) and the SEPs are detected with anti-FLAG antibody (green). ASNSD1-SEP, PHF19-SEP, and EIF5-SEP are all derived from sORFs in the 5'-UTR (uORFs); cells expressing EIF5 are indicated with a white arrow. Two additional 5'-UTR sORFs, FRAT2-SEP and YTHD3-SEP, produce SEPs initiating with a non-AUG codons. Finally, DEDD2-SEP (CDS) and H2AFx-SEP (3'-UTR) were not translated from the RefSeq RNAs, which is consistent with a scanning model of eukaryotic translation. (B) Diagrams of the SEP-encoding RNAs used to produce the results depicted in A (the diagrams for ASNSD1 and DEDD2 are shown in Figure 1.3A). The gray arrow represents the RNA transcript, the blue arrow represents annotated protein CDS and the yellow arrow represents the sORF. The numbers mark the boundaries of each element of the transcript in nucleotide bases from the 5' end of the transcript.

A

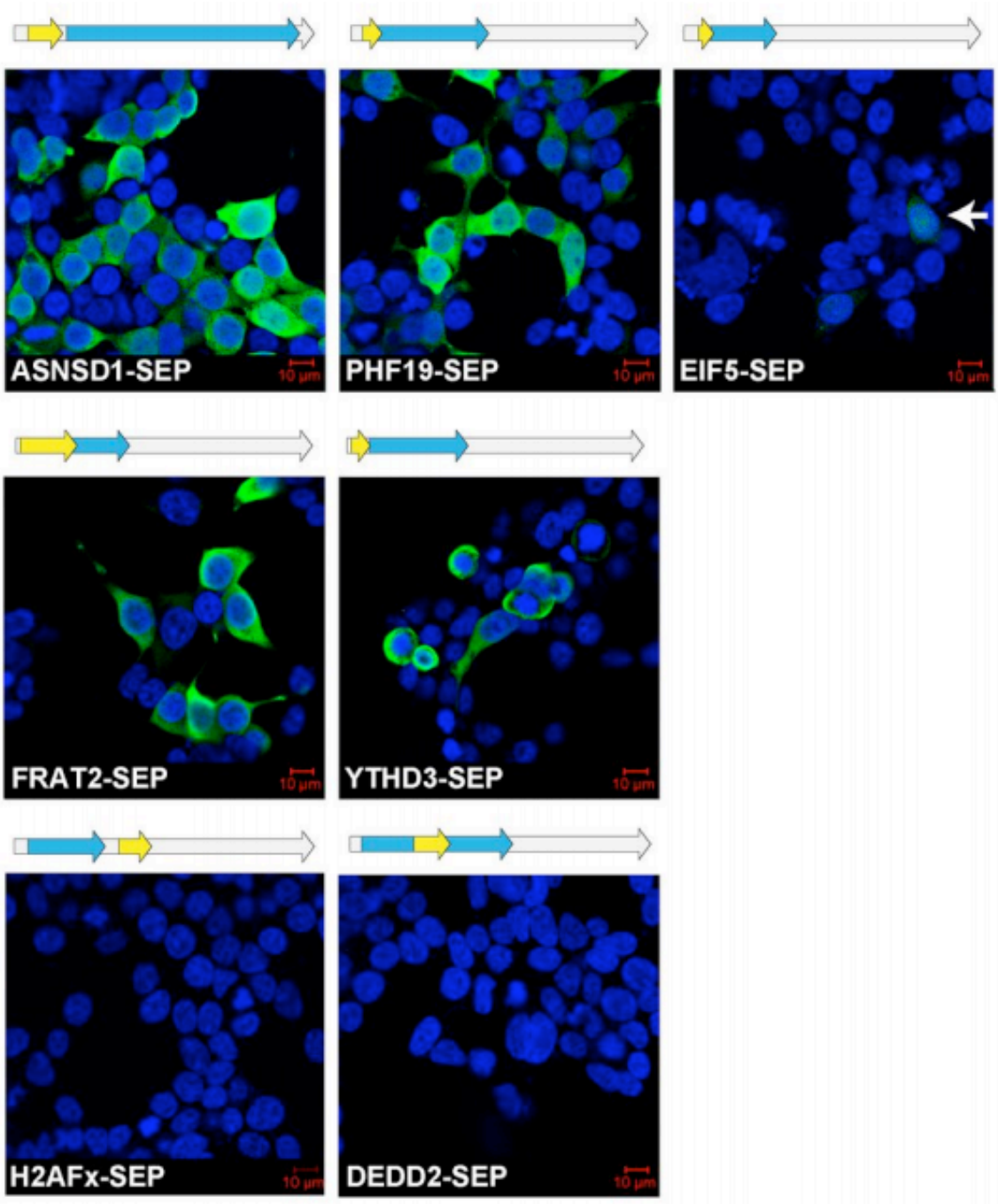


Figure 1.7 (continued)

B

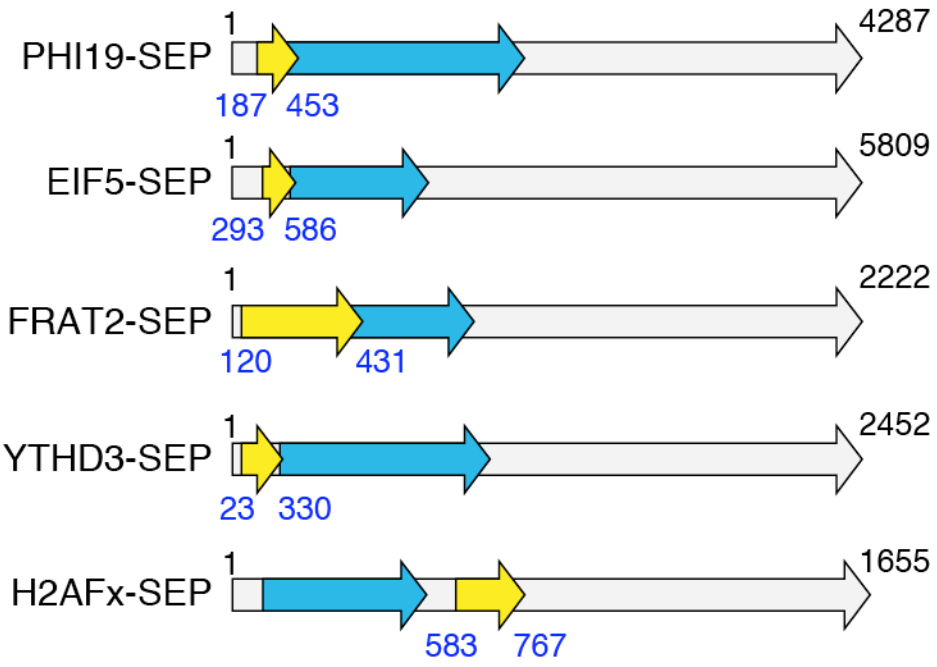


Figure 1.7 (continued)

produced SEPs verifies that sORFs with non-AUG start codons are translated (Figure 1.7A).

By contrast, the *DEDD2-SEP* sORF was not translated from the full-length RefSeq construct (Figure 1.7A). *DEDD2-SEP* is frameshifted deep within the main CDS of the DEDD2 transcript, so according to the scanning model of translation (57) it is not expected that this downstream sORF would be translated (Figure 1.7B). Similarly, the 3'-UTR-embedded H2AFx-SEP was similarly not translated from the full-length mRNA construct (Figure 1.7A). One possible explanation for these observations is that the DEDD2-SEP and the H2AFx-SEP are translated from splice variants of their respective annotated transcripts that are present in K562 cells but are not in the RefSeq database. In any case, it would seem that not all SEPs that derive from a segment of an annotated transcript are translated from a bicistronic mRNA.

1.11 Alternative splicing of annotated protein coding genes generates SEP-producing transcripts

We identified a truncated DEDD2 mRNA in the RNA-seq data wherein the first start codon is that of the *DEDD2-SEP* sORF, making the transcript ideal for translation of DEDD2-SEP by the traditional ribosome scanning mechanism (Figure 1.8). This supports the hypothesis that alternative splicing of annotated protein coding genes is one mechanism by which SEP-producing transcripts are produced. However, we were not able to clearly identify a truncated version of the

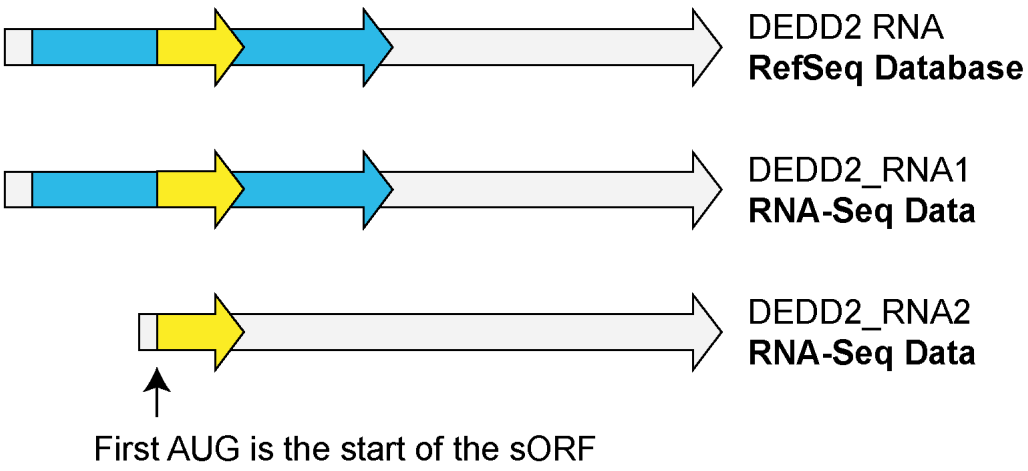


Figure 1.8 Alternative splicing of SEP-encoding transcripts. The expression of the DEDD2-SEP in K562 cells may be explained by the existence of an alternative splice form of the DEDD2 RNA (DEDD2_RNA2), the first start codon of which initiates *DEDD2-SEP*. In this figure, the gray arrow represents the RNA, the blue arrow represents annotated protein CDS (if present), and the yellow arrow represents the sORF.

H2AFx transcript in the K562 RNA-seq data. It is possible that a truncated H2AFx mRNA variant is present in K562 cells but did not give rise to the sequencing reads necessary to resolve it from the full-length H2AFx transcript.

1.12 Confirming a non-AUG start site and investigating the mechanism of bicistronic expression

Since such a large proportion of SEPs putatively initiate at non-AUG sites, we wanted to rigorously identify the alternate start codon of one these sORFs. C-terminally FLAG-tagged FRAT2-SEP was expressed from the full-length mRNA construct in HEK293T cells and immunoprecipitated; mass spectrometry of the purified protein (Figure 1.9) was consistent with initiation at an ACG triplet embedded within a Kozak consensus sequence (38) (Figure 1.10A). Mutating the ACG to an ATG resulted in increased FRAT2-SEP translation while deletion of this ACG abolished FRAT2-SEP production, as assessed by Western blotting, thus confirming our assignment (Figure 1.10B). In addition, mutation of the Kozak consensus sequence to less favorable residues led to markedly lower FRAT2-SEP expression, which demonstrates the importance of the Kozak sequence at non-AUG initiation sites.

The scanning model of translation provided an explanation as to why the DEDD2 mRNA is not bicistronic; we hypothesized that upstream alternate start codons could provide a mechanism to promote polycistronic gene expression via leaky scanning. To test whether FRAT2 mRNA is bi-cistronic, we prepared a

FRAT2 construct where the SEP and the downstream CDS were tagged with different epitopes (Figure 1.10C), permitting their simultaneous detection by immunoblotting with two antibodies. We found that the FRAT2 RNA is bi-cistronic, as FRAT2 and FRAT2-SEP are both expressed (Figure 1.10C).

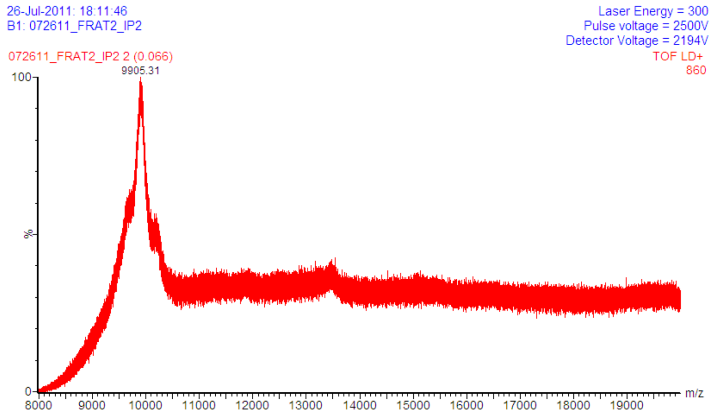


Figure 1.9 Identifying the start site of FRAT2-SEP. MALDI-MS of immunoprecipitated FRAT2-SEP-FLAG provides a polypeptide with a molecular weight of 9905, which corresponds to an ACG initiation codon with methionine as the first amino acid.

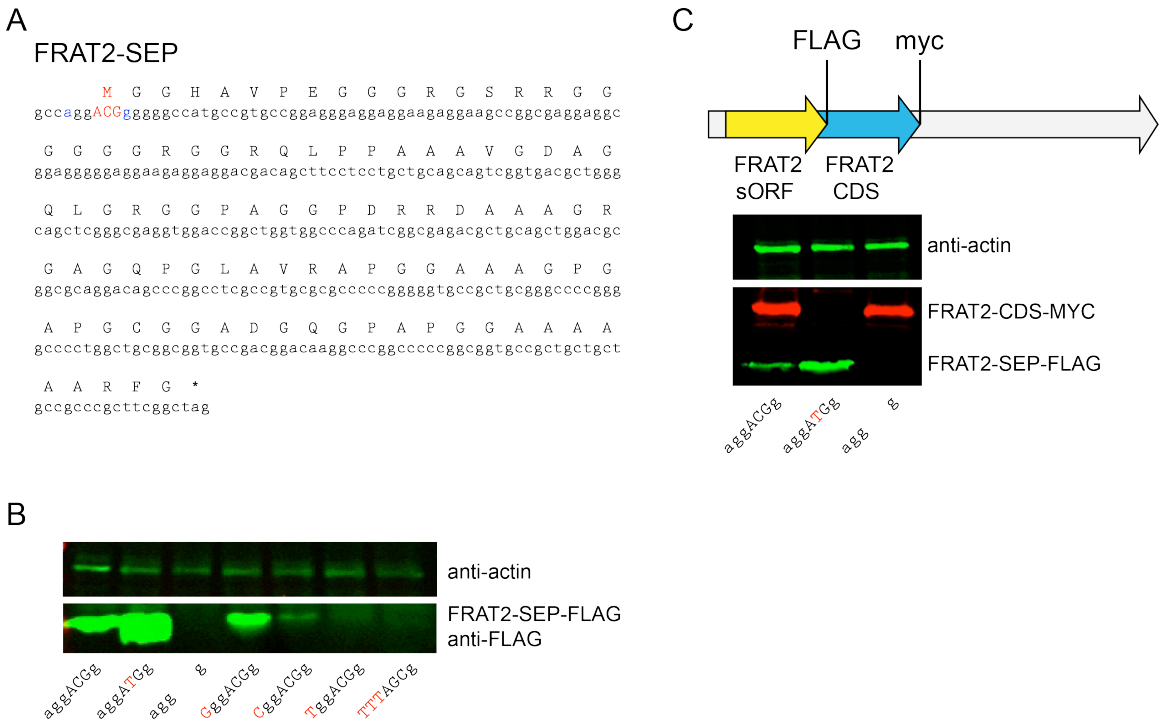
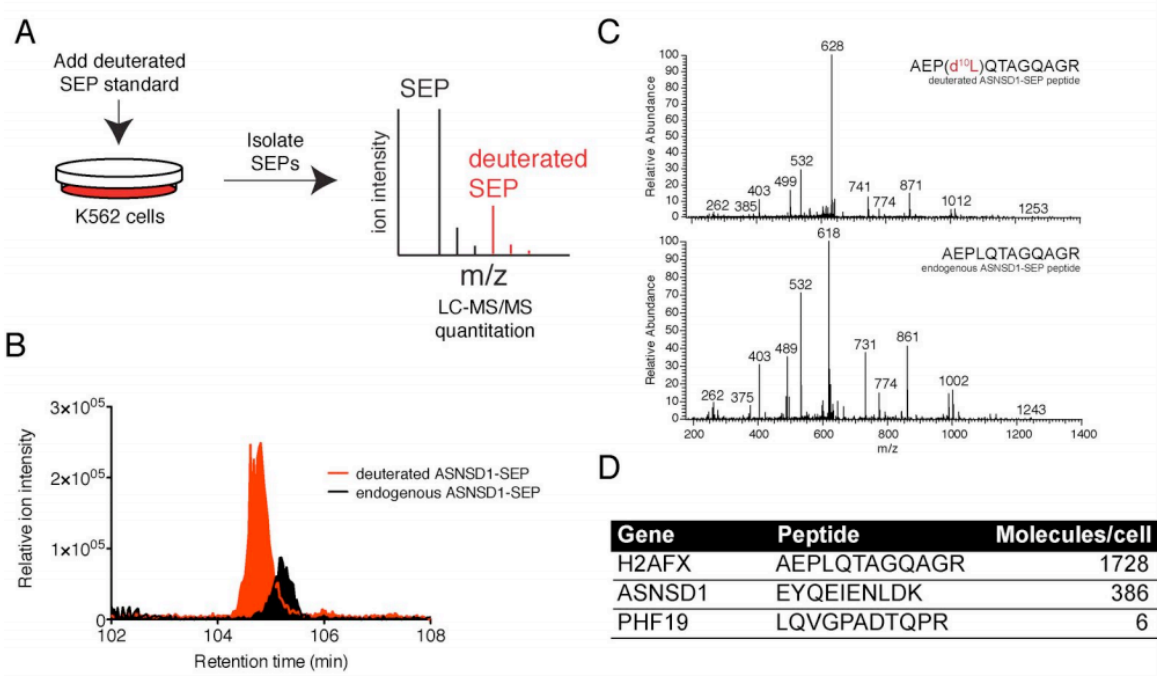


Figure 1.10 Characterization of the non-AUG initiation codon of the *FRAT2-SEP* sORF. (A) An ACG triplet embedded in a Kozak consensus sequence was identified as the *FRAT2-SEP* initiation codon (red) by determining the molecular weight of immunoprecipitated FRAT2-SEP-FLAG using MALDI-MS. (B) This ACG was confirmed as the *FRAT2-SEP* initiation codon by site-directed mutagenesis followed by western blots of FRAT2-SEP-FLAG using an anti-FLAG antibody. Conversion of the ACG to an ATG resulted in higher expression (lane 2), while ablation of this codon removed all expression (lane 3). In addition, perturbation of the Kozak sequence (lanes 4-7) revealed the importance of context when using non-AUG codons, as substitution of less favorable residues (38) at the most important positions in the Kozak sequence resulted in lower FRAT2-SEP-FLAG expression. (C) Epitope tagging of the sORF and CDS of the FRAT2 mRNA demonstrates that the FRAT2 mRNA is bi-cistronic. Specifically, the FRAT2 CDS was *c*-myc tagged and the FRAT2-SEP was FLAG tagged. Conversion of the FRAT2-SEP initiation codon from ACG to ATG ablates the expression of the downstream FRAT2-CDS, indicating the importance of alternate start codons for polycistronic expression.

Remarkably, mutation of the ACG start codon of the *FRAT2-SEP* to an ATG increases FRAT2-SEP expression, but also completely eliminates the expression of FRAT2 protein, revealing that the translation of the downstream cistron absolutely requires leaky upstream initiation. Therefore, this experiment indicated that an upstream non-AUG initiation codon is necessary for efficient polycistronic gene expression.

1.13 Measuring the cellular concentrations of SEPs

As a first step towards exploring the functional potential of the discovered SEPs we wished to determine whether they persist in the cell as concentrations that are comparable to that of known functional peptides and proteins. We therefore measured the cellular concentrations (K562 cells) of three randomly selected SEPs (ASNSD1-SEP, PHF19-SEP and H2AFx-SEP) using isotope dilution mass spectrometry (58) (Figure 1.11). In these experiments, isotopically heavy-labeled peptides corresponding to the detected peptides that were used to identify these three SEPs were synthesized and added to the cells during extraction. The sample was then processed as described above (Figure 1.1) except that the ERLIC fractionation step was omitted. Removal of the ERLIC speeds up the sample preparation but also results in less sensitivity during the LC-MS analysis. Therefore, we compensated for the loss of ERLIC by using selected ion monitoring (SIM) during the LC-MS, which is more sensitive. Analysis of the lysate by LC-MS is able to distinguish the light (endogenous) and heavy



Gene	Peptide	Molecules/cell
H2AFX	AEPLQTAGQAGR	1728
ASNSD1	EYQEIENLDK	386
PHF19	LQVGPADTQPR	6

Figure 1.11 SEP quantification. (A) SEPs were quantified by isotope dilution-mass spectrometry (IDMS). We synthesized a deuterated (heavy-labeled) variant of the diagnostic SEP peptide we detected. Upon isolation of K562 cells this peptide was added and the entire mixture was separated using our standard approach to isolate SEPs. SEPs are then quantified by LC-MS since the deuterated SEP and endogenous SEP can be distinguished by differences in their masses. (B) Overlap between the endogenous SEP and the deuterated SEP along with (C) matching MS/MS spectra (note: 10 Da shift for heavy peptide for some fragments), indicate that these are the same peptides. (D) Quantification of several SEPs using this approach.

(exogenously added) variants of the SEP, and since a known amount of heavy labeled SEP was added, the ratio of light-to-heavy can be used to quantify the absolute amount of endogenous peptide. These SEPs were found to have concentrations of between 10 and 2000 copies per cell (Figure 1.11D). Thus, based on previous estimates of protein copy numbers, SEPs are found at concentrations well within the range of that of typical cellular proteins (59-61). We further note that the MS/MS spectra from the synthetic standards used in these experiments were nearly identical to those produced from the endogenous peptide and eluted at the same retention time as same, thus confirming these identifications (Figure 1.11C).

1.14 SEPs exhibit sub-cellular localization

After failing to observe expression of DEDD2-SEP from the full-length DEDD2 transcript (Figure 1.7), we subcloned an expression construct for FLAG-tagged DEDD2-SEP to determine whether the peptide could be stably expressed. Interestingly, we found that DEDD2-SEP localizes to mitochondria in HEK293T, mouse embryonic fibroblast (MEF), and COS7 cells, as demonstrated by co-localization with the mitochondrial marker MitoTracker Red (Figure 1.13A-C). This finding suggests that DEDD2-SEP may have a role in mitochondrial function, which is interesting because the DEDD2 CDS has been implicated in apoptosis (62, 63), which is largely dependent on the mitochondria. The N-terminus of DEDD2-SEP is predicted to contain a mitochondrial import signal

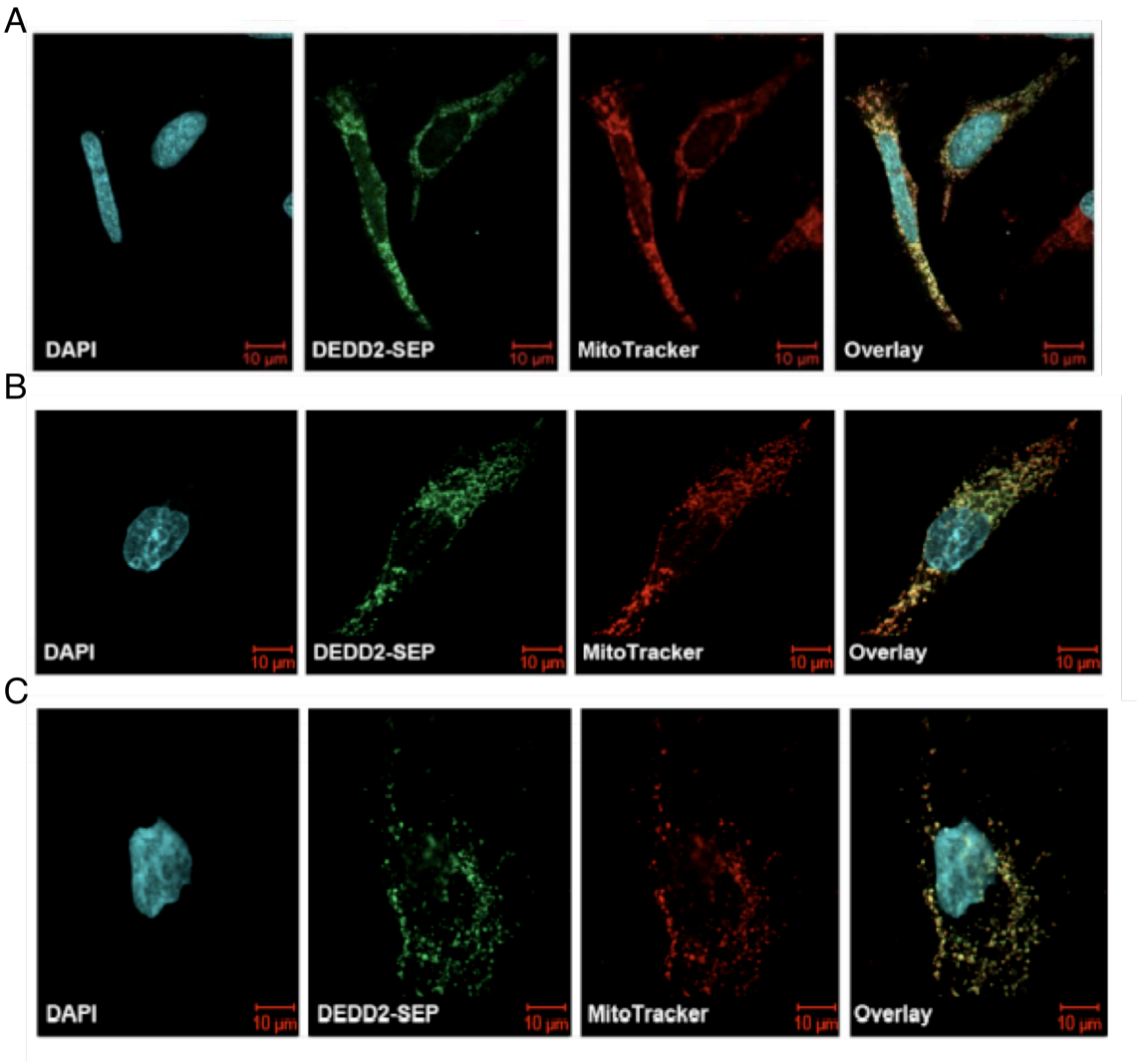


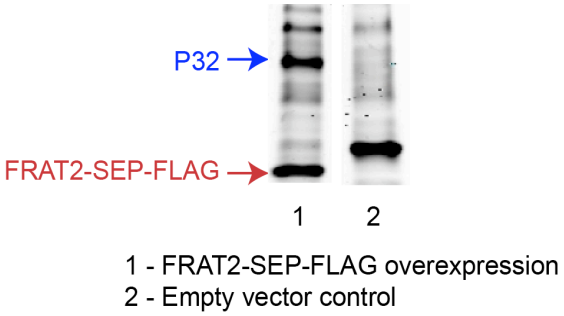
Figure 1.12 DEDD2-SEP localizes to the mitochondria. DEDD2-SEP was subcloned and expressed in HeLa (A), MEF (B) and COS7 (C) cells to examine its expression and localization by immunofluorescence. Co-staining with MitoTracker (red) indicated that the DEDD2-SEP localizes to the mitochondria (overlay).

(64). Sequence-dependent trafficking and sub-cellular localization of SEPs could therefore be general phenomena related to their biological activities.

1.15 SEPs participate in protein-protein interactions

During the analysis of FRAT2-SEP expressed in HEK293T cells (Figure 1.6), we observed reproducible co-immunoprecipitation of a higher molecular weight protein (Figure 1.14A). This signaled the presence of a potential protein-protein interaction partner. Proteomic analysis of this protein band revealed the putative FRAT2-SEP binding partner as mitochondrial P32 (P32, HABP or C1QBP) (65). We confirmed that this protein-protein interaction occurs in the parent K562 cell line (and therefore may occur endogenously) by immunoprecipitation of transiently transfected FLAG-tagged FRAT2-SEP. We detected co-immunoprecipitation of P32 in transfected cells by immunoblotting with an anti-P32 antibody, with no background in cells transfected with empty vector (Figure 1.14B). The fact that FRAT2-SEP specifically interacts with another protein is highly suggestive that SEPs will be found to have cellular functions.

A



B

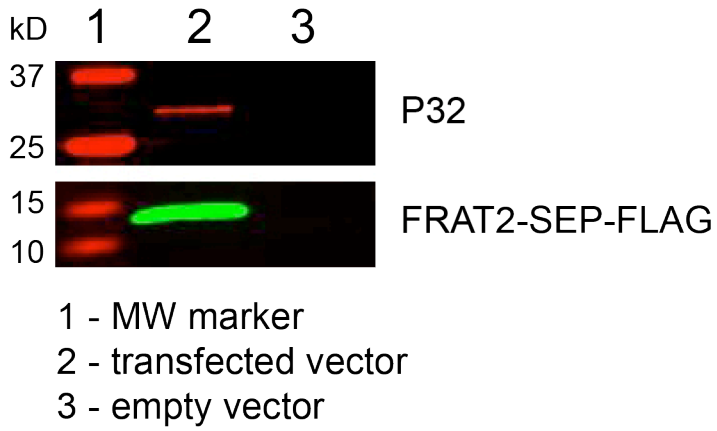


Figure 1.14 FRAT2-SEP participates in a protein-protein interaction with P32. (A) A Krypton fluorescent protein stained gel showing a band co-immunoprecipitating with FRAT2-SEP-FLAG. This band migrates at approximately 30 kD. Excision of this band and proteomic analysis revealed its identity as mitochondrial P32 (P32). (B) The FRAT2-SEP-P32 interaction was confirmed in K562 cells by transient transfection and immunoprecipitation of FRAT2-SEP-FLAG followed by immunoblotting with anti-P32 antibody.

1.16 SEPs influence gene expression

The serendipitous discoveries that SEPs participate in protein-protein interactions and exhibit subcellular localization piqued our interest in the functional potential of these molecules. We wondered whether some SEPs possess other bioactivities, such as the ability to influence gene expression. We therefore performed gene expression profiling experiments on 10 SEPs.

The microarray is the gold standard in global gene expression assays; however, this technique is time consuming and expensive. Recently, a group led by Aravind Subramanian of the Broad Institute of MIT and Harvard developed a high-throughput and highly cost-effective alternative for measuring gene expression on a global scale. The method is based on the observation that gene expression is highly correlated. By determining the expression levels of a carefully selected set of “landmark” genes, this group has found, it is possible to accurately infer the expression profile of the remainder of the genome using known correlation patterns. Specifically, the assay involves determining the mRNA levels of 1000 genes using a ligation-mediated amplification (LMA) protocol coupled with Luminex bead technology (66). In the LMA step, amplification is accomplished using locus-specific probes containing unique molecular barcodes and universal biotinilated primers. Amplified transcripts are incubated in streptavidin-phycoetherin, hybridized to optically-addressed beads that anneal to specific “barcode” sequences, and analyzed using a two-laser flow cytometer, whereby one laser detects the bead color (denoting transcript identity)

and the other laser detects phycoetherin fluorescence (indicating transcript abundance).

To assess the effects of SEPs on gene expression, we used the L1000 assay to compare the gene expression profiles of two sets of samples: an experimental set in which cells (HEK293T) were transfected with an overexpression construct containing the sORF encoding the SEP to be studied and a control set in which cells were transfected with an overexpression construct containing the sORF encoding that SEP but with the start codon ablated so no polypeptide would be produced. For each SEP, cells were harvested at three time points: 10 hours, 20 hours and 30 hours. Overexpression was verified by immunofluoresence imaging (all SEPs were C-terminally FLAG-tagged). These samples (n=6) were then analyzed by the L1000 assay to determine the differential expression pattern of the experimental sample.

Six out of the 10 SEPs produced statistically significant signatures (p-value < 0.05). Of the six SEPs tested, two – C7ORF49-SEP and DNLZ-SEP – appeared to produce significant changes in gene expression, as indicated by a heat map illustrating the expression differential for the 50 most up-regulated and the 50 most down-regulated genes (Figure 1.15). While this result is significant because it indicates that these SEPs have biological activity, we wanted to get more insight in the nature of the signatures our SEPs produced. For this, we turned to a knowledge-based method for analyzing differential gene expression data called Gene Set Enrichment Analysis (GSEA) (67).

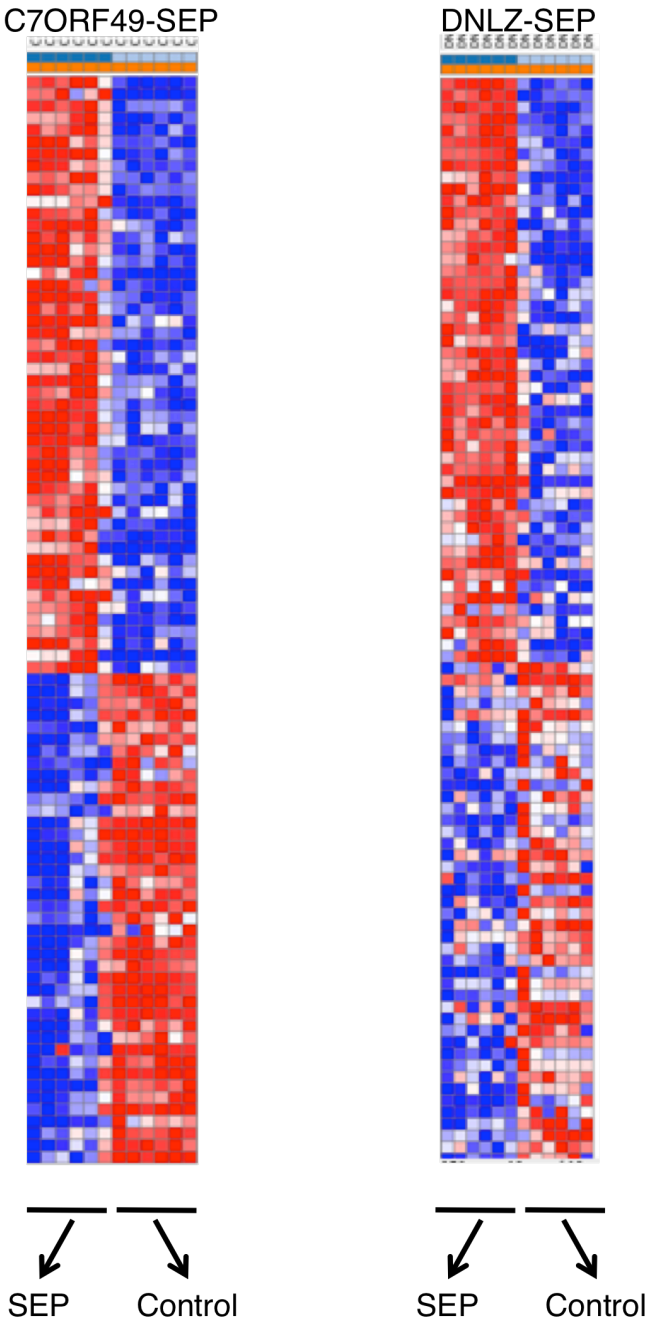


Figure 1.15 Heat map of the expression levels of the 50 most up-regulated genes and the 50 most down-regulated genes in the C7ORF49-SEP and DNLZ-SEP overexpression experiments. Gene expression levels in HEK293T cells in which SEP was overexpressed are shown on the left of each diagram. Gene expression levels in the control samples (no SEP) are shown on the right of each diagram. Red indicates high expression and blue indicates low expression.

In GSEA, genes are sorted into a rank-ordered list according to their differential expression levels and then a set of genes that is known to be involved in a particular pathway is mapped onto the list. The degree to which that gene set is enriched in the sample – that is, the degree to which it is overrepresented at the top or bottom of the list – is assessed by 'walking' down the list and increasing a running-sum statistic when a gene in the set is encountered and decreased the metric when a gene that is not in the list is encountered. The resulting metric is called the enrichment score (ES). When normalized for gene set size, the ES can be used to determine whether the perturbagen used in the experiment (in this case, a SEP) modulates the gene expression of this set to a degree that would indicate the perturbagen targets a regulation pathway that controls the expression of the gene set. An NES score whose absolute value is greater than 2.25 indicates a substantial enrichment of the gene set in question.

The GSEA analysis of the expression signature produced by DNLZ-SEP indicated that this SEP targets the same gene sets as heat shock factor protein 2 (HSF2), estrogen receptor beta 2 (ESR2) and transcription factor ETV6 (Table 1.4). Interestingly, chromosomal aberrations involving ETV6 are found in chronic myelomonocytic leukemia (CMML) and acute myeloid leukemia (AML). It is therefore conceiveable that this SEP plays a role in producing the disease phenotype of the cells used in this study, which are myelogenous leukemia cells. C7ORF49-SEP also produced a signature that was enriched with curated gene sets, several of which are involved in controlling basic processes in translation

(Table 1.5). These results demonstrate that SEPs are capable of modulating gene expression on a global scale, and in some cases in a manner similar to that of known functional species.

Table 1.4 Results of gene set enrichment analysis for DNLS-SEP. The normalized enrichment score (NES) indicated the degree to which the SEP up-regulated genes in the set. A score above 2.25 indicates substantial up-regulation. Detailed descriptions of the gene sets can be found at http://www.broadinstitute.org/cmap/.

Gene Set	NES	Nominal p-value	FDR
HF2506_UP_desc:HSF2	2.61	0	0
HF1775_UP_desc:ESR2	2.29	0	0
HF2164_UP_desc:ETV6	2.29	0	0

Table 1.5 Results of gene set enrichment analysis for C7ORF49-SEP. The normalized enrichment score (NES) indicated the degree to which the SEP up-regulated genes in the set. A score above 2.25 indicates substantial up-regulation. Detailed descriptions of the gene sets can be found at http://www.broadinstitute.org/cmap/.

Gene Set	NES	Nominal p-value	FDR
REACTOME_60S_RIBOSOMAL_SUBUNIT	5.009	0	0
REACTOME_FORMATION_OF_40S_SUBUNITS	4.890	0	0
HSIAO_HOUSEKEEPING_GENES	4.797	0	0
KEGG_RIBOSOME	4.564	0	0
KEGG_SPLICEOSOME	3.995	0	0.02

1.17 Conclusion

In conclusion, we used a novel approach to discover the largest number of human SEPs ever reported. Importantly, we successfully access the pool of SEPs that are under 50 amino acids in length, including two peptides that are a mere 14 residues long. This is unprecedented for mass spectrometry-based discovery approach and is a crucial step towards understanding the biology of these molecules; for as we learned from the example of *polished rice* in Drosophila (12), SEPs as small as 11 amino acids can play a significant functional roles in animals. Moreover, the smallest SEPs are those least likely to be discovered by comparative genomics approaches due the distinctive challenge of detecting the evolutionary signature of conservation over short protein-coding sequences.

We also uncovered several unexpected features of SEPs, among them that SEP translation is frequently initiated at non-AUG codons and that SEPs can arise from polycistronic mRNAs. Highlighting the interplay between these features, we find that the non-AUG start codon of one of our SEPs is necessary for efficient polycistronic gene expression. This may explain why such a large fraction of the discovered SEPs initiate at non-AUG sites. Lastly, we determined that SEPs persist in the cell at concentrations that are comparable to known functional proteins. Perhaps most intriguingly, though, we find that SEPs possess properties characteristic of functional proteins, such as stable expression, high cellular copy numbers, post-translational modifications, subcellular localization,

the ability to participate in specific protein-protein interactions and the ability to influence gene expression.

Taken together, these findings indicate that the human proteome is significantly more complex than previously appreciated. Moreover, due to the bias of mass spectrometry for more abundant species (55), which limits the scope of our technique to the most highly expressed or most stable SEPs, it is probable that there are many more as-yet-undiscovered human SEPs. Thus, we believe we have only begun to explore the breadth and diversity of this exciting new family of polypeptides.

1.18 Materials and methods

Cloning and mutagenesis

DNA constructs were prepared by standard ligation, Quikchange, or inverse PCR techniques. Human cDNA clones were obtained from Open Biosystems. Gene synthesis was by DNA2.0. Plasmid sequences are publicly available at http://web.me.com/saghatelian/Saghatelian_Lab/Home.html or upon request. We note that the YTHDF3-SEP construct consisted of the 5'-UTR putatively encoding the SEP only, obtained via gene synthesis because a full-length cDNA construct with an intact 5'-UTR was not commercially available.

Cell culture

Cells were grown at 37ºC under an atmosphere of 5% CO_2. HEK293T, HeLa, COS7 and MEF cells were grown in high-glucose DMEM supplemented with L-glutamine, 10% fetal bovine serum, penicillin and streptomycin. K562 cells were maintained at a density of 1-10 x 10^5 cells/mL in RPMI1640 media with 10% fetal bovine serum, penicillin and streptomycin.

Isolation and processing of polypeptides

Aliquots of 3 x 10^7 growing K562 cells were placed in 1.5 ml Protein LoBind Tubes (Eppendorf), washed three times with PBS, pelleted and stored at -80 °C. Boiling water (500 µl) was added directly to the frozen cell pellets and the samples were then boiled for 20 minutes to eliminate proteolytic activity (20, 21). After cooling to room temperature, samples were sonicated on ice for 20 bursts at output level 4 with a 40% duty cycle (Branson Sonifier 250; Ultrasonic Convertor). The cell lysate was then brought to 0.25% acetic acid by volume and centrifuged at 20,000 x g for 20 minutes at 4°C. The supernatant was sent through a 30 kD or 10 kD molecular weight cut-off (MWCO) filter (Modified PES Centrifugal Filter, VWR). The mix of small proteins and peptides in the flow-through was evaluated for protein content by BSA assay and then evaporated to dryness at low temperature in a SpeedVac. Pellets were re-suspended in 50 µl of 25mM TCEP in 50mM NH_4HCO_3 (pH=8) and incubated at 37 °C for 1 hour. The reaction was cooled to room temperature before 50 µl of a 50 mM iodoacetamide

solution in 50 mM NH_4HCO_3. This solution was incubated in the dark for 1 hour. Samples were then dissolved in a 50 mM NH_4HCO_3 solution of 20 µg/µl trypsin (Promega) to a final protein to enzyme mass ratio of 50:1. This reaction was incubated at 37 °C for 16 hours, cooled to room temperature and then quenched by adding neat formic acid to 5% by volume. The digested peptide mix was then bound to a C18 Sep Pak cartridge (HLB, 1cm^3; 30mg, Oasis), washed thoroughly with water and eluted with 1:1 acetonitrile/water. The eluate was evaporated to dryness at low temperature on a SpeedVac.

Offline electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) fractionation of polypeptide fraction

To simplify the sample and thereby improve detection sensitivity in the subsequent LC-MS/MS analysis, we separated the processed peptide mix by ERLIC (28, 29). ERLIC was performed using a PolyWax LP column (200 x 2.1 mm, 5µm, 300Å; PolyLC Inc.) connected to an Agilent Technologies 1200 Series HPLC equipped with a degasser and automatic fraction collector. All runs were performed at a flow rate of 0.3 ml/min and ultraviolet absorption was measured at a wavelength of 210 nm. Forty (30 kD sample) or 25 (10 kD sample) fractions were collected over a 70 minute gradient beginning with 0.1% acetic acid in 90% acetonitrile (aq.) and ending with 0.1% formic acid in 30% acetonitrile (aq.). The fractions were then evaporated to dryness on a SpeedVac and dissolved in 15 µl 0.1% formic acid (aq.) in preparation for LC-MS/MS analysis.

LC-MS/MS analysis. Samples were injected onto a NanoAcquity HPLC system (Waters) equipped with a 5 cm x 100 µm capillary trapping column (New Objective) packed with 5 um C18 AQUA beads (Waters) and a PicoFrit SELF/P analytical column (15 µm tip, 25 cm length, New Objective) packed with 3 µm C18 AQUA beads (Waters) and separated over a 90 minute gradient at 200 nl/min. This HPLC system was online with an LTQ Orbitrap Velos (Thermo Scientific) instrument, which collected full MS (dynamic exclusion) and tandem MS (Top 20) data over 375-1600 m/z with 60,000 resolving power.

Data processing

The acquired MS/MS spectra were analyzed with the SEQUEST algorithm using a database derived from 6-frame (forward and reverse) translation of RefSeq (National Center for Biotechnology Information) mRNA transcripts or 3-frame (forward only) translation of a transcriptome assembly generated by Cufflinks (Trapnell et al 2010) using RNA-Seq data from the K562 cell line (data acquisition described below). The search was performed with the following parameters: variable modifications, oxidation (Met), N-acetylation; semitryptic requirement; maximum missed cleavages: 2; precursor mass tolerance: 20 ppm; and fragment mass tolerance: 0.7 Da. Search results were filtered such that the estimated false discovery rate of the remaining results was 1%. The Sf score is the final score for protein identification by the Proteomics Browser software based on a combination of the preliminary score, the cross-correlation and the

differential between the scores for the highest scoring protein and second highest scoring protein (34).

Identified peptides were searched against the Uniprot human protein database using a string-searching algorithm. Peptides found to be identical to fragments of annotated proteins were eliminated from the SEP candidate pool. The remaining peptides were searched against non-redundant protein sequences using the Basic Local Alignment Search Tool (BLAST). Any peptides found to be less than two amino acids different from the nearest protein match (i.e., identical or different by one amino acid) were discarded.

The spectra of the remaining peptides were subjected to a rigorous manual validation procedure: spectra were rejected if they had a precursor mass error of >5 ppm, if they lacked a sequence tag of 5 consecutive b- or y-ions, if they had more than one missed cleavage, or if they lacked sufficient sequence coverage to differentiate from the nearest annotated protein.

RNA-Seq library preparation, alignment, and transcriptome assembly

Two types of cDNA libraries were generated from K-562 RNA (Ambion). In the first experiment, we used 50 nanograms of polyA$^+$ RNA to create standard, non-strand-specific cDNA libraries with paired-end adaptors as previously described (68) and sequenced it on one lane of an Illumina Genome Analyzer IIa machine. In the second experiment, we used eight different amounts of total RNA (30 ng, 100 ng, 250 ng, 500 ng, 1000ng, 3000 ng, and 10,000 ng) to create cDNA

libraries with paired-end, indexed adaptors following the instructions for the Illumina TruSeq RNA sample prep kit, except that we used SuperScript III instead of SuperScript II and optimized PCR cycle number. These libraries were sequenced on a single lane of a HiSeq2000 machine. RNA-Seq reads were aligned to the human genome (Hg19 assembly) using TopHat [version V1.1.4; Trapnell et al. Bioinformatics 2009] and transcriptome assembly was performed using Cufflinks [version V1.0.0; (69)]. lincRNAs were called based on a lincRNA-calling pipeline as previously described (70). The transcriptome data is deposited on GEO (GSE34740).

Peptide synthesis, purification and concentration determination

Automated (PS3 Protein Technology, Inc.) solid-phase peptide synthesis was carried out using Fmoc amino acids. Crude peptides were HPLC (Shimadzu)-purified using a C18 column (150 mm × 20 mm, 10 μm particle size, Higgins Analytical). The mobile phase was adjusted for each peptide; buffer A was 99% H_2O, 1% acetonitrile, and 0.1% TFA; buffer B was 90% acetonitrile, 10% H_2O, and 0.07% TFA). Pure fractions were identified by MALDI-MS analysis, combined, and lyophilized. Peptide concentrations were determined by amino acid analysis (AlBio Tech).

Absolute quantification of SEPs

Isotope dilution mass spectrometry (IDMS) (58) was used to determine the concentration of SEPs in K562 cells. All samples for this experiment were prepared by adding known amounts of heavy isotope-labeled peptides corresponding the detected fragment of the SEP of interest to a K562 cell pellet (10^7 cells) just before isolation of the polypeptides from these cells. The preparation of these samples was identical to that described above except that no ERLIC separation was done. The first step of the quantification procedure was to prepare a set of samples where each sample contained a different but known amount (1 fmol, 10 fmol, 50 fmol, 100 fmol, 500 fmol, 1 pmol or 10 pmol) of the heavy-labeled counterpart peptide. These samples were then analyzed by a selected ion monitoring (SIM) method on the previously described LC-MS/MS system and the resulting data was analyzed using Xcaliber 2.0 (Thermo Scientific). The areas of the peaks corresponding to the endogenous and isotope-labeled peptides were compared to determine the approximate concentration of the SEP and a standard curve was generated to verify that the quantity of the SEP fragment was within the linear range of the mass spectrometer. A second set of samples that each contained an amount of isotope-labeled peptide that was within the linear range of the instrument and within an order of magnitude of the amount of the corresponding endogenous peptide in the cells was then prepared (N=4) and analyzed as described. The

results of this experiment were used to determine the absolute cellular concentration of the selected SEPs.

Imaging SEPs by immunofluorescence

HeLa, COS7, and MEF cells were grown to 80% confluency on glass coverslips in 48-well plates; HEK293T cells were grown to 50-75% confluency on fibronectin (Millipore)-coated glass coverslips in 48-well plates. Cells were transfected in Opti-MEM (Invitrogen) with 250 ng plasmid DNA using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. 24 hours after transfection, cells were fixed with 4% formalin in phosphate buffered saline (PBS) for 10 minutes at room temperature, and then permeabilized with methanol at -20ºC for 10 minutes. Fixed cells were blocked with blocking buffer (3% BSA in PBS with 0.5% Tween-20), then incubated overnight at 4ºC with anti-FLAG M2 antibody (Sigma) diluted 1:1000 in blocking buffer. After washing 3x with PBS, cells were then stained for one hour at room temperature with goat anti-mouse AlexaFluor 488 conjugate (Invitrogen) diluted 1:1000 in blocking buffer. Cells were washed 3x with PBS, post-fixed with 4% formalin for 10 minutes at room temperature, then counterstained with a final concentration of 270 ng/mL Hoescht 33258 (Invitrogen) in PBS for 15 minutes at room temperature. Cells were then imaged in PBS in glass-bottom imaging dishes (Matek Corp.). For mitochondrial co-localization analysis, transfected cells were treated with MitoTracker Red CMXRos (Invitrogen) at a final concentration of 100 nM in PBS

at 37ºC for 15 minutes, washed once with PBS, then fixed with formalin and methanol and immunostained as described above.

Images were acquired in the Harvard Center for Biological Imaging on a Zeiss LSM 510 inverted confocal microscope with the following lasers: 405 Diode, 488 (458,477,514) Argon, 543 HeNe and 633 HeNe. Image acquisition was with either AIM or Zen software. Images were acquired with a 60x oil immersion objective.

Determination of the FRAT2-SEP start codon by immunoprecipitation and MALDI-MS

COS7 and HEK293T cells were grown in 10-cm dishes to 75% confluency, then transfected with 10 μg plasmid DNA using Lipofectamine 2000 according to the manufacturer's instructions. 24 hours after transfection, cells were harvested by scraping and washed 3x with PBS. Cells were lysed in 400 μL Triton lysis buffer (1% Triton X-100 in Tris-buffered saline (TBS) with Roche Complete Mini Protease Inhibitor added) on ice for 15 minutes, then lysates were clarified by centrifugation at 16,100 x g for 20 minutes at 4ºC. Clarified lysates were added to 50 μL of PBS-washed anti-FLAG M2 agarose resin (Sigma) and rotated at 4ºC for 1 hour. Beads were washed 6x with TBS-T (Tris-buffered saline with 0.05% Tween-20). To elute bound proteins, 50 uL of 100 ug/mL 3x FLAG peptide (Sigma) in TBS-T was added to the resin and rotated at 4ºC for 20 minutes. Eluates were stored at -80ºC until further analysis.

For MALDI-MS analysis, the entire protein sample was desalted using a C18 Sep Pak cartridge (HLB, 1cm^3; 30mg, Oasis) and eluted in 50% acetonitrile. The sample was dried in a SpeedVac, and then dissolved in a final volume of 10 μL mass spectrometry-grade water (Burdick & Jackson). This solution (1 μL) was mixed with matrix (α-cyano-4-hydroxycinnamic acid in 50% acetonitrile, 1 μL) on a stainless steel MALDI plate and air-dried. Data were acquired on a Waters MALDI micro MX instrument operated in linear positive mode. Instrument control and spectral acquisition were with MassLynx software.

Confirmation of the FRAT2-SEP initiation codon, Kozak sequence, and bicistronic expression by immunoblotting

HEK293T cells were grown to 75% confluency in 6-well plates, then transfected with 10 μg plasmid DNA using Lipofectamine 2000 according to the manufacturer's instructions. Cells were harvested by vigorous pipetting and lysed in 100 μL Triton lysis buffer. Samples of clarified lysate (20 μL) were mixed with SDS-PAGE loading buffer, boiled, and electrophoresed on 4-20% Tris-HCl gels (Bio-Rad). Two replicate gels were run. Proteins were transferred to nitrocellulose (0.20 μm pore size, Thermo Scientific) and immunoblots were probed with anti-FLAG M2 antibody (Sigma) followed by goat anti-mouse IR dye 800 conjugate (LICOR). For bicistronic expression assays, immunoblots were probed with a mixture of rabbit anti-c-myc antibody (Sigma) and anti-FLAG M2, followed by a mixture of goat anti-mouse IR dye 800 and goat anti-rabbit IR dye

680 (LICOR). A replica immunoblot was probed with mouse anti-β-actin followed by goat anti-mouse IR dye 800. Antibodies were diluted 1:2000 in Rockland Immunochemicals fluorescent blocking buffer. Infrared imaging was performed on a LICOR Odyssey instrument.

Co-immunoprecipitation, LC-MS proteomics, and immunoblotting to detect FRAT2 protein-protein interaction partners

HEK293T cells were grown in 10-cm dishes to 75% confluency, then transfected with 10 μg plasmid DNA using Lipofectamine 2000 according to the manufacturer's instructions. Cells were harvested, lysed, and subjected to immunoprecipitation as described.

For SDS-PAGE analysis, 20 uL of the eluate was boiled in SDS-PAGE loading buffer, then electrophoresed on a 4-20% Tris-HCl gel (Bio-Rad). Gels were stained with Coommassie brilliant blue and bands of interest were excised with a clean razor blade and subjected to in-gel trypsin digest and LC-MS/MS analysis. Proteins were identified using Sequest.

To confirm P32 binding, HEK293T cells were grown and transfected as above, then co-immunoprecipitation with anti-FLAG agarose was performed as described. A sample (10 µL) of each eluate was boiled in SDS-PAGE loading buffer, then electrophoresed on a 4-20% Tris-HCl gel. Three replicate gels were run. Proteins were transferred to nitrocellulose and probed with rabbit anti-C1QBP antibody (Sigma) followed by goat anti-rabbit IR dye 680 conjugate. A

replica immunoblot was probed with anti-FLAG M2 antibody followed by goat anti-mouse IR dye 800 conjugate. Antibodies were diluted 1:2000 in Rockland Immunochemicals fluorescent blocking buffer. A third replica gel was stained with Krypton protein stain (Pierce) to detect total protein loading.

To confirm P32 binding to FRAT2-SEP in K562 cells, cells were passaged to a density of 5 x 10^5 cells/mL in Opti-MEM (Invitrogen) in 10-cm dishes and grown for 24 hours. Once cells reached a density of 1 x 10^6 cells/mL, 10 μg of plasmid DNA was transfected with Lipofectamine according to the manufacturer's instructions. Cells were harvested 24 hours after transfection by centrifuging for 5 minutes at 100 x g, washed 3x with PBS, and lysed as described. Co-immunoprecipitation and immunoblotting were performed as described.

Preparing samples for Luminex 1000 analysis

HEK293T cells at 75% confluence in 96-well plate wells were transfected as described in "Imaging SEPs by immunofluorescence" (n=6). Cells were also transfected in 48-well plates in parallel to serve as tests of transfection efficiency for each construct employed. At the relevant time point (10, 20 or 30 hours after transfection), cells were gently washed 3x with PBS and 100 ul of Buffer TCL (Qiagen Inc.) was added. The 48-well plate samples were then prepared and imaged as described in "Imaging SEPs by immunofluorescence". Plates were then covered in aluminum foil and gently shaken for 30 minutes to ensure

complete lysis. Finally, plates were placed in a -80 C freezer for storage until Luminex 1000 analysis could be performed.

1.19 References

1. Calvo SE, Pagliarini DJ, & Mootha VK (2009) Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc Natl Acad Sci U S A* 106(18):7507-7512.

2. Parola AL & Kobilka BK (1994) The peptide product of a 5' leader cistron in the beta 2 adrenergic receptor mRNA inhibits receptor synthesis. *J Biol Chem* 269(6):4497-4505.

3. Werner M, Feller A, & Messenguy F (1987) The leader peptide of yeast gene CPA1 is essential for the translational repression of its expression. *Cell*.

4. Wadler CS & Vanderpool CK (2007) A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. *Proc Natl Acad Sci U S A* 104(51):20454-20459.

5. Jay G, Nomura S, Anderson CW, & Khoury G (1981) Identification of the SV40 agnogene product: a DNA binding protein.

6. Casson SA*, et al.* (2002) The POLARIS gene of Arabidopsis encodes a predicted peptide required for correct root growth and leaf vascular patterning. *Plant Cell* 14(8):1705-1721.

7. Rohrig H, Schmidt J, Miklashevichs E, Schell J, & John M (2002) Soybean ENOD40 encodes two peptides that bind to sucrose synthase. *Proc Natl Acad Sci U S A* 99(4):1915-1920.

8. Kastenmayer JP*, et al.* (2006) Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. *Genome Res* 16(3):365-373.

9. Gleason CA, Liu QL, & Williamson VM (2008) Silencing a candidate nematode effector gene corresponding to the tomato resistance gene Mi-1 leads to acquisition of virulence. *Mol Plant Microbe Interact* 21(5):576-585.

10. Galindo MI, Pueyo JI, Fouix S, Bishop SA, & Couso JP (2007) Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol* 5(5):e106.

11. Kondo T*, et al.* (2007) Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nat Cell Biol* 9(6):660-665.

12. Hashimoto Y*, et al.* (2001) A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer's disease genes and Abeta. *Proc Natl Acad Sci U S A* 98(11):6336-6341.

13. Frith MC*, et al.* (2006) The abundance of short proteins in the mammalian proteome. *PLoS Genet* 2(4):e52.

14. Zhang F & Hinnebusch AG (2011) An upstream ORF with non-AUG start codon is translated in vivo but dispensable for translational control of GCN4 mRNA. *Nucleic Acids Res* 39(8):3128-3140.

15. Hemm MR, Paul BJ, Schneider TD, Storz G, & Rudd KE (2008) Small membrane proteins found by comparative genomics and ribosome binding site models. *Molecular Microbiology* 70(6):1487-1501.

16. Oyama M*, et al.* (2007) Diversity of translation start sites may define increased complexity of the human short ORFeome. *Mol Cell Proteomics* 6(6):1000-1006.

17. Ingolia NT, Lareau LF, & Weissman JS (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147(4):789-802.

18. Svensson M, Skold K, Svenningsson P, & Andren PE (2003) Peptidomics-based discovery of novel neuropeptides. *J Proteome Res* 2(2):213-219.

19. Hummon AB*, et al.* (2006) From the genome to the proteome: uncovering peptides in the Apis brain. *Science* 314(5799):647-649.

20. Tagore DM*, et al.* (2009) Peptidase substrates via global peptide profiling. *Nat Chem Biol* 5(1):23-25.

21. Tinoco AD, Tagore DM, & Saghatelian A (2010) Expanding the dipeptidyl peptidase 4-regulated peptidome via an optimized peptidomics platform. *J Am Chem Soc* 132(11):3819-3830.

22. Lozzio CB & Lozzio BB (1975) Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. *Blood* 45(3):321-334.

23. Svensson M, Skold K, Svenningsson P, & Andren PE (2003) Peptidomics-based discovery of novel neuropeptides. *J Proteome Res* 2(2):213-219.

24. Che FY, Lim J, Pan H, Biswas R, & Fricker LD (2005) Quantitative neuropeptidomics of microwave-irradiated mouse brain and pituitary. *Molecular & cellular proteomics : MCP* 4(9):1391-1405.

25. Parkin MC, Wei H, O'Callaghan JP, & Kennedy RT (2005) Sample-dependent effects on the neuropeptidome detected in rat brain tissue preparations by capillary liquid chromatography with tandem mass spectrometry. *Analytical chemistry* 77(19):6331-6338.

26. Skold K*, et al.* (2007) The significance of biochemical and molecular sample integrity in brain proteomics and peptidomics: stathmin 2-20 and peptides as sample quality indicators. *Proteomics* 7(24):4445-4456.

27. Van Dijck A*, et al.* (2011) Comparison of extraction methods for peptidomics analysis of mouse brain tissue. *Journal of neuroscience methods* 50(3):227-232.

28. Alpert AJ (2008) Electrostatic repulsion hydrophilic interaction chromatography for isocratic separation of charged solutes and selective isolation of phosphopeptides. *Anal Chem* 80(1):62-76.

29. Hao P*, et al.* (2010) Novel application of electrostatic repulsion-hydrophilic interaction chromatography (ERLIC) in shotgun proteomics: comprehensive profiling of rat kidney proteome. *J Proteome Res* 9(7):3520-3526.

30. Pruitt KD, Tatusova T, & Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 33(Database issue):D501-504.

31. Yassour M*, et al.* (2010) Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species. *Genome Biol* 11(8):R87.

32. Eng JK, McCormack AL, & Yates Iii JR (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry* 5(11):976-989.

33. Yates JR, 3rd, Eng JK, McCormack AL, & Schieltz D (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal Chem* 67(8):1426-1436.

34. Christofk HR, Vander Heiden MG, Wu N, Asara JM, & Cantley LC (2008) Pyruvate kinase M2 is a phosphotyrosine-binding protein. *Nature* 452(7184):181-186.

35. Li M*, et al.* (2011) Widespread RNA and DNA sequence differences in the human transcriptome. *Science* 333(6038):53-58.

36. Trapnell C*, et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28(5):511-515.

37. Altschul SF, Gish W, Miller W, Myers EW, & Lipman DJ (1990) Basic local alignment search tool. *Journal of molecular biology* 215(3):403-410.

38. Kozak M (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* 44(2):283-292.

39. Okazaki Y*, et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420(6915):563-573.

40. Zhao J, Sun BK, Erwin JA, Song JJ, & Lee JT (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* 322(5902):750-756.

41. Heo JB & Sung S (2011) Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science* 331(6013):76-79.

42. Rinn JL*, et al.* (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129(7):1311-1323.

43. Leighton PA, Ingram RS, Eggenschwiler J, Efstratiadis A, & Tilghman SM (1995) Disruption of imprinting caused by deletion of the H19 gene region in mice. *Nature* 375(6526):34-39.

44. Pandey RR*, et al.* (2008) Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Molecular cell* 32(2):232-246.

45. Vanderpool CK & Gottesman S (2007) The novel transcription factor SgrR coordinates the response to glucose-phosphate stress. *Journal of bacteriology* 189(6):2238-2248.

46. Rice JB & Vanderpool CK (2011) The small RNA SgrS controls sugar-phosphate accumulation by regulating multiple PTS genes. *Nucleic Acids Res* 39(9):3806-3819.

47. Maki K, Morita T, Otaka H, & Aiba H (2010) A minimal base-pairing region of a bacterial small RNA SgrS required for translational repression of ptsG mRNA. *Molecular microbiology* 76(3):782-792.

48. Kawamoto H, Morita T, Shimizu A, Inada T, & Aiba H (2005) Implication of membrane localization of target mRNA in the action of a small RNA: mechanism of post-transcriptional regulation of glucose transporter in Escherichia coli. *Genes & development* 19(3):328-338.

49. Lin MF, Jungreis I, & Kellis M (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* 27(13):i275-282.

50. Cabili MN*, et al.* (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development* 25(18):1915-1927.

51. Punta M*, et al.* (2012) The Pfam protein families database. *Nucleic Acids Res* 40(Database issue):D290-301.

52. Guttman M*, et al.* (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458(7235):223-227.

53. Guttman M*, et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* 28(5):503-510.

54. Khalil AM*, et al.* (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A* 106(28):11667-11672.

55. Fonslow BR*, et al.* (2011) Improvements in proteomic metrics of low abundance proteins through proteome equalization using ProteoMiner prior to MudPIT. *J Proteome Res* 10(8):3690-3700.

56. Pruitt KD, Tatusova T, & Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 35(Database issue):D61-65.

57. Hinnebusch AG (2011) Molecular mechanism of scanning and start codon selection in eukaryotes. *Microbiol Mol Biol Rev* 75(3):434-467, first page of table of contents.

58. Keshishian H, Addona T, Burgess M, Kuhn E, & Carr SA (2007) Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution. *Mol Cell Proteomics* 6(12):2212-2229.

59. de Godoy LM*, et al.* (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature* 455(7217):1251-1254.

60. Schwanhausser B*, et al.* (2011) Global quantification of mammalian gene expression control. *Nature* 473(7347):337-342.

61. Beck M*, et al.* (2011) The quantitative proteome of a human cell line. *Mol Syst Biol* 7:549.

62. Roth W, Stenner-Liewen F, Pawlowski K, Godzik A, & Reed JC (2002) Identification and characterization of DEDD2, a death effector domain-containing protein. *J Biol Chem* 277(9):7501-7508.

63. Alcivar A, Hu S, Tang J, & Yang X (2003) DEDD and DEDD2 associate with caspase-8/10 and signal cell death. *Oncogene* 22(2):291-297.

64. Bendtsen JD, Nielsen H, von Heijne G, & Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 340(4):783-795.

65. Itahana K & Zhang Y (2008) Mitochondrial p32 is a critical mediator of ARF-induced apoptosis. *Cancer Cell* 13(6):542-553.

66. Peck D*, et al.* (2006) A method for high-throughput gene expression signature analysis. *Genome biology* 7(7):R61.

67. Subramanian A*, et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102(43):15545-15550.

68. Levin JZ*, et al.* (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods* 7(9):709-715.

69. Trapnell C*, et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28(5):511-515.

70. Cabili MN*, et al.* (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 25(18):1915-1927.

Chapter 2

Substance P: A Case Study in the Challenges of Investigating Bioactive Peptide Regulation

2.1 Introduction

Peptides constitute a major class of signaling molecules in animals and have been shown to play a role in diverse physiological processes, including glucose homeostasis (insulin) (1), hypertension (angiotensin) (2), social behavior (oxytocin) (3), appetite (ghrelin) (4) and sleep (orexins) (5-9). As a result, elucidating the molecular mechanisms by which bioactive peptides are regulated is an interesting and important research endeavor.

Protease cleavage is a well-established paradigm for the regulation of peptide levels and signaling *in vivo*, and pharmacologically targeting proteolytic pathways that control peptide levels has proven effective in treating human disease. For example, angiotensin converting enzyme (ACE) inhibitors, which prevent the production of the vasoconstrictor peptide angiotensin II, are commonly used to treat hypertension (10). Similarly, dipeptidyl peptidase 4 (DPP4) inhibitors, which raise insulin levels by preventing the degradation of the insulinotropic hormone glucagon-like peptide 1 (GLP1), are a new class of anti-diabetes drugs (11). However, despite the clear basic and biomedical importance of understanding the proteolytic pathways controlling peptide levels *in vivo*, these processes have been defined for only a small fraction of the known bioactive peptides.

In this chapter, I explore the strategies that have historically been used to probe bioactive peptide regulation by examining the case of substance P, which was the first bioactive peptide discovered and is probably the most well studied.

Understanding the strengths and pitfalls of these methods suggests a path towards more efficient and effective studies. In the concluding section, I outline a promising new approach that integrates analytical chemistry, classical biochemistry and genetic or pharmacological studies in animal models into a general platform for defining the biochemical pathways that control the activity of bioactive peptides. This powerful method lacks many of the deficiencies of traditional approaches and in theory can be readily applied to any bioactive peptide. In chapter 5, I discuss a study that brings this method to bear on substance P.

2.2 Substance P and its functional role in mammals

Discovered in 1931, substance P is a bioactive undecapeptide (Figure 2.1) that is widely expressed in the central and peripheral nervous systems (12) of mammals and functions as a neurotransmitter and neuromodulator (13). It has been shown to play a role in such diverse processes as pain transmission (14-16), inflammation (17, 18), sleep (19), learning and memory (19-21), depression and affective mood disorders (22-24), opioid dependence (25-27) and apoptosis (28, 29).

Arg^{1}-Pro-Lys-Pro-Gln^{5}-Gln-Phe-Phe-Gly-Leu^{10}-Met-NH_2

Figure 2.1 The structure of substance P. The mature peptide is amidated at the C-terminal.

2.3 Substance P biogenesis, secretion and mechanism of action

A member of the tachykinin family of neuropeptides, substance P is synthesized from the preprotachykinin-A gene into a large, biologically inert protein precursor, or prepropeptide (29). Converting enzymes process the prepropeptide to release a shorter precursor peptide, which is then packaged into vesicles (30, 31) and axonally transported to terminal endings (32). There, proteases and amidating enzymes produce the mature, active form of substance P, which can be secreted into the synapse to perform its function as a neurotransmitter and neuromodulator (32).

Mechanistically, substance P elicits its effects by binding to one of three G-protein coupled receptors that are specific to the tachykinin peptide family: neurokinin 1 (NK_1), neurokinin 2 (NK_2) or neurokinin 3 (NK_3) (33-37). However, it has the greatest affinity for NK_1 and this is believed to be the primary endogenous receptor for the peptide (38, 39). Radiolabelling and mutatgenesis studies indicate that the key facet of substance P- NK_1 coupling is insertion of the hydrophobic sequence at the carboxy-terminus of substance P (GLM-NH_2) into a hydrophobic binding pocket on NK_1 (40). The fact that SP(6-11) is the smallest fragment of the peptide that retains significant affinity for NK_1 underscores the importance of the C-terminal region for effective binding (41). Upon stimulation by substance P, NK_1 transmits signals into the cell through one or more of three second-messenger systems: phosphatidyl inositol stimulation via phospholipase

C, which leads to Ca^{2+} influx (42-44); arachidonic acid release via phospholipase A2 (43, 45); or cAMP mobilization via adenylate cyclase(44, 46, 47).

2.4 Substance P inactivation

Although there is evidence that some fraction of substance P is degraded within the endosome after the clathrin-mediated endocytosis that can occur after substance P- NK_1 binding (48), the majority of secreted substance P is not taken back up by the cell but rather is inactivated or converted to other active forms by proteolytic cleavage in the intercellular space (49). Several enzymes have been suggested to participate in substance P metabolism. Dipeptidyl aminopeptidase IV (DPP4) and prolyl endopeptidase (PREP), for example, have been shown to degrade Substance P from the N-terminus *in vitro* (50-52). These enzymes produce the C-terminal fragments SP(3-11) and SP(5-11). Additional evidence supporting PREP as a Substance P-regulating enzyme is the finding that PREP inhibition slightly increases substance P levels in the brain (53, 54). DPP4, on the other hand, is virtually absent from the brain, making it an unlikely candidate for regulating substance P in the central nervous system.

Other enzymes have been found to cleave substance P at the C-terminus or exhibit an endopeptidase activity. Angiotensin converting enzyme (ACE), which co-localizes with substance P in some regions of the brain (55-58), releases substance P fragments SP(1-7) and SP(1-8) when incubated with full-length peptide (56). Neutral endopeptidase 24.11 (NEP), another co-localizing

enzyme (59), cleaves the peptide after Gln^6, Phe^7 and Gly^9 (60, 61). Further implicating ACE and NEP in substance P metabolism is the fact that inhibitors of these enzymes suppress substance P-degrading activity in brain tissue lysates (62, 63). However, more recent studies have revealed that some potent inhibitors of ACE and NEP fail to protect substance P from degradation, indicating the previous results were due to off-target effects and making it unlikely that these enzymes participate in endogenous substance P metabolism (64). Moreover, none of the aforementioned enzymes displays a strong specificity for substance P and, notably, most are known to act on a variety of other neuropeptides (65).

In addition to these well-characterized enzymes, several peptidases exhibiting a high specificity for substance P have been reported. One such enzyme, isolated from the membrane fraction of human brain, hydrolyzed substance P at the Gln^6-Phe^7, Phe^7-Phe^8 and Phe^8-Gly^9 bonds and was described as a neutral metalloendopeptidase with a molecular weight of 40-50kDa (66). An approximately 70kD enzyme possessing a very similar activity and specificity for substance P was isolated from rat substance Pinal cord and given the name substance P-degrading enzyme (SPDE) (66, 67). Another substance P-specific enzyme was found in human cerebrospinal fluid and given the name substance P endopeptidase (SPE) (52). It cleaved substance P at the Phe^7-Phe^8 and Phe^8-Gly^9 bonds. Subsequently, enzymes closely resembling SPE in affinity for substance P, cleavage pattern and inhibitor sensitivity were purified from human spinal cord (68), rat spinal cord (69) and rat brain ventral

tegemental area (70). However, the genes from which these substance P-specific enzymes derive remain unknown, making rigorous study of their impact on endogenous substance P metabolism difficult.

2.5 *In vitro* approaches to studying substance P degradation

Ultimately, identifying the enzymes responsible for degrading substance P may depend on determining which fragments of the peptide are actually products of endogenous metabolism; certainly, such information will accelerate the identification process. To this end, many researchers have conducted *in vitro* experiments where cell lysates are incubated with synthetic peptide (52, 71-73) or pseudo-*in vivo* experiments where synthetic peptide is injected into tissues (64, 74-76). In these studies, a multitude of fragments have been detected: SP(1-2), SP(1-4), SP(1-7), SP(1-8), SP(1–9), SP(1-10), SP(4–8), SP(6–10), SP(7–10), SP(3–11), SP(5–11), SP(6–11), SP(7–11), SP(8–11). Strong evidence suggests a physiological role for SP(1-7) (26, 77-80), establishing the peptide as a likely product of endogenous substance P metabolism. Of the remaining fragments, some, for example SP(1-4), SP(1-8), SP(1-9) and SP(5-11), have also been reported to possess biological activity (15, 81-83), though it is not clear whether these have a natural function or whether their activity is simply 'accidental' due to chemical similarity with the parent peptide.

These fragments and others could be the final products of inactivation cuts, intermediates in conversion pathways, or perhaps independently functioning

species. It is also possible, however, that many of the detected fragments result from non-specific peptidase activity present in the homogenate or injected tissue but which does not participate in natural substance P-degrading pathways. Indeed, whenever exogenous peptide is used in a metabolic study, one cannot be sure the results are reflective of endogenous metabolism. Additionally, there is the confounding factor of location. Substance P is expressed extensively in the nervous system, and several of its biological functions seem to be organ-specific (84). Thus, insights gleaned from examination of the hypothalamus may not be relevant to metabolism at nerve endings in the gut, and results from whole brain homogenates or CSF may conflate the metabolism of multiple tissues. In light of these considerations, I think it is apparent that a fast, effective approach to elucidating the regulation of substance P must include *in vivo* measurements of peptide levels in the tissue of interest and seek to draw candidate peptidases from the pool of enzymes known to be present and active in that tissue.

2.6 *In vivo* approaches to studying substance P degradation

Overall, the approaches taken to studying substance P metabolism reflect the approaches taken to studying bioactive peptide regulation in general. Much work aimed at identifying the metabolites of bioactive peptides has relied on *in vitro* experiments, which have the limitations discussed above. Some researchers, though, have used *in vivo* peptide measurements to inform their efforts, and this has led to successful identification of the acting peptidases. For example, the

discovery of DPP4 as a GLP-1 degrading enzyme (85) relied on an approach that combined HPLC fractionation with radioimmunoassay (RIA) to detect endogenous metabolites of GLP-1 in blood plasma (86) (87).

Despite the evident effectiveness of RIA and other immunoassays (e.g., ELISA, EIA) at detecting small quantities of peptide, these methods have some significant drawbacks. The potential for antibody cross-reactivity, for example, means these assays are not necessarily selective, and the possibility that antibodies will not sufficiently react with all possible metabolites – or not react at all with metabolites that have undergone modification (i.e., oxidation) – means these assays are not necessarily thorough (76, 88). Furthermore, the need for specialized reagents (antibodies) makes these methods difficult to incorporate into an efficient, general approach to studying bioactive peptide metabolism.

2.7 Conclusion

In contrast to the techniques discussed above, a mass spectrometry-based peptidomics strategy (53, 89-92) could provide detailed information on the *in vivo* metabolism of a bioactive peptide without the disadvantages of immunoassays and, importantly, could be readily applied to any bioactive peptide of interest. Coupling this technique with *in vitro* biochemical assays, an enzyme purification/identification workflow and a genetic strategy to test for biological relevance would create a powerful platform for proteolytic pathway discovery. Indeed, such an approach has already been used to characterize the metabolism

of peptide histidine isolucine (PHI), a bioactive peptide that promotes glucose-stimulated insulin secretion, in the intestine (93). In Chapter 5, I describe a study in which a customized version of this method was applied to substance P regulation in the spinal cord.

2.8 References

1. Saltiel AR & Kahn CR (2001) Insulin signalling and the regulation of glucose and lipid metabolism. *Nature* 414(6865):799-806.

2. Patchett AA*, et al.* (1980) A new class of angiotensin-converting enzyme inhibitors. *Nature* 288(5788):280-283.

3. Jin D*, et al.* (2007) CD38 is critical for social behaviour by regulating oxytocin secretion. *Nature* 446(7131):41-45.

4. Nakazato M*, et al.* (2001) A role for ghrelin in the central regulation of feeding. *Nature* 409(6817):194-198.

5. Peyron C*, et al.* (2000) A mutation in a case of early onset narcolepsy and a generalized absence of hypocretin peptides in human narcoleptic brains. *Nature medicine* 6(9):991-997.

6. Thannickal TC*, et al.* (2000) Reduced number of hypocretin neurons in human narcolepsy. *Neuron* 27(3):469-474.

7. Chemelli RM*, et al.* (1999) Narcolepsy in orexin knockout mice: molecular genetics of sleep regulation. *Cell* 98(4):437-451.

8. Lin L*, et al.* (1999) The sleep disorder canine narcolepsy is caused by a mutation in the hypocretin (orexin) receptor 2 gene. *Cell* 98(3):365-376.

9. Hara J*, et al.* (2001) Genetic ablation of orexin neurons in mice results in narcolepsy, hypophagia, and obesity. *Neuron* 30(2):345-354.

10. Atkinson AB & Robertson JI (1979) Captopril in the treatment of clinical hypertension and cardiac failure. *Lancet* 2(8147):836-839.

11. Herman GA, Stein PP, Thornberry NA, & Wagner JA (2007) Dipeptidyl peptidase-4 inhibitors for the treatment of type 2 diabetes: focus on sitagliptin. *Clin Pharmacol Ther* 81(5):761-767.

12. Ribeiro-da-Silva A & Hokfelt T (2000) Neuroanatomical localisation of Substance P in the CNS and sensory neurons. *Neuropeptides* 34(5):256-271.

13. Otsuka M & Yoshioka K (1993) Neurotransmitter functions of mammalian tachykinins. *Physiol Rev* 73(2):229-308.

14. Hall ME & Stewart JM (1983) Substance P and antinociception. *Peptides* 4(1):31-35.

15. Hall ME & Stewart JM (1983) Substance P and behavior: opposite effects of N-terminal and C-terminal fragments. *Peptides* 4(5):763-768.

16. Yeomans DC & Proudfit HK (1992) Antinociception induced by microinjection of substance P into the A7 catecholamine cell group in the rat. *Neuroscience* 49(3):681-691.

17. Kolasinski SL, Haines KA, Siegel EL, Cronstein BN, & Abramson SB (1992) Neuropeptides and inflammation. A somatostatin analog as a selective antagonist of neutrophil activation by substance P. *Arthritis Rheum* 35(4):369-375.

18. Palmer JM & Greenwood B (1993) Regional content of enteric substance P and vasoactive intestinal peptide during intestinal inflammation in the parasitized ferret. *Neuropeptides* 25(2):95-103.

19. Schlesinger K*, et al.* (1986) Substance P facilitation of memory: effects in an appetitively motivated learning task. *Behav Neural Biol* 45(2):230-239.

20. Nagel JA, Welzl H, Battig K, & Huston JP (1993) Facilitation of tunnel maze performance by systemic injection of the neurokinin substance P. *Peptides* 14(1):85-95.

21. Santangelo EM, Morato S, & Mattioli R (2001) Facilitatory effect of substance P on learning and memory in the inhibitory avoidance test for goldfish. *Neurosci Lett* 303(2):137-139.

22. Berrettini WH*, et al.* (1985) CSF substance P immunoreactivity in affective disorders. *Biol Psychiatry* 20(9):965-970.

23. Shirayama Y, Mitsushio H, Takashima M, Ichikawa H, & Takahashi K (1996) Reduction of substance P after chronic antidepressants treatment in the striatum, substantia nigra and amygdala of the rat. *Brain Res* 739(1-2):70-78.

24. Lieb K, Treffurth Y, Berger M, & Fiebich BL (2002) Substance P and affective disorders: new treatment opportunities by neurokinin 1 receptor antagonists? *Neuropsychobiology* 45 Suppl 1:2-6.

25. Nylander I, Sakurada T, Le Greves P, & Terenius L (1991) Levels of dynorphin peptides, substance P and CGRP in the spinal cord after subchronic administration of morphine in the rat. *Neuropharmacology* 30(11):1219-1223.

26. Kreeger JS & Larson AA (1993) Substance P-(1-7), a substance P metabolite, inhibits withdrawal jumping in morphine-dependent mice. *Eur J Pharmacol* 238(1):111-115.

27. Kreeger JS & Larson AA (1996) The substance P amino-terminal metabolite substance P(1-7), administered peripherally, prevents the development of acute morphine tolerance and attenuates the expression of withdrawal in mice. *J Pharmacol Exp Ther* 279(2):662-667.

28. Kang BN*, et al.* (2001) Regulation of apoptosis by somatostatin and substance P in peritoneal macrophages. *Regul Pept* 101(1-3):43-49.

29. Lallemend F*, et al.* (2003) Substance P protects spiral ganglion neurons from apoptosis via PKC-Ca2+-MAPK/ERK pathways. *J Neurochem* 87(2):508-521.

30. Merighi A*, et al.* (1988) Ultrastructural studies on calcitonin gene-related peptide-, tachykinins- and somatostatin-immunoreactive neurones in rat dorsal root ganglia: evidence for the colocalization of different peptides in single secretory granules. *Cell Tissue Res* 254(1):101-109.

31. Plenderleith MB, Haller CJ, & Snow PJ (1990) Peptide coexistence in axon terminals within the superficial dorsal horn of the rat spinal cord. *Synapse* 6(4):344-350.

32. Brimijoin S, Lundberg JM, Brodin E, Hokfelt T, & Nilsson G (1980) Axonal transport of substance P in the vagus and sciatic nerves of the guinea pig. *Brain Res* 191(2):443-457.

33. Gerard NP*, et al.* (1991) Human substance P receptor (NK-1): organization of the gene, chromosome localization, and functional expression of cDNA clones. *Biochemistry* 30(44):10640-10646.

34. Teichberg VI, Cohen S, & Blumberg S (1981) Distinct classes of substance P receptors revealed by a comparison of the activities of substance P and some of its segments. *Regul Pept* 1(5):327-333.

35. Lee CM, Iversen LL, Hanley MR, & Sandberg BE (1982) The possible existence of multiple receptors for substance P. *Naunyn-Schmiedeberg's archives of pharmacology* 318(4):281-287.

36. Buck SH, van Giersbergen PL, & Burcher E (1991) Tachykinins and their receptors: Pharmacology, biochemistry and molecular biology advance a neuropeptide story to the forefront of science. *Neurochemistry international* 18(2):167-170.

37. Regoli D, Drapeau G, Dion S, & D'Orleans-Juste P (1987) Pharmacological receptors for substance P and neurokinins. *Life Sci* 40(2):109-117.

38. Nakanishi S (1991) Mammalian tachykinin receptors. *Annual review of neuroscience* 14:123-136.

39. Regoli D, Boudon A, & Fauchere JL (1994) Receptors and antagonists for substance P and related peptides. *Pharmacol Rev* 46(4):551-599.

40. Huang RR, Yu H, Strader CD, & Fong TM (1994) Interaction of substance P with the second and seventh transmembrane domains of the neurokinin-1 receptor. *Biochemistry* 33(10):3007-3013.

41. Lee CM, Campbell NJ, Williams BJ, & Iversen LL (1986) Multiple tachykinin binding sites in peripheral tissues and in brain. *Eur J Pharmacol* 130(3):209-217.

42. Mochizuki-Oda N, Nakajima Y, Nakanishi S, & Ito S (1994) Characterization of the substance P receptor-mediated calcium influx in cDNA transfected Chinese hamster ovary cells. A possible role of inositol 1,4,5-trisphosphate in calcium influx. *J Biol Chem* 269(13):9651-9658.

43. Nakajima Y, Tsuchida K, Negishi M, Ito S, & Nakanishi S (1992) Direct linkage of three tachykinin receptors to stimulation of both phosphatidylinositol hydrolysis and cyclic AMP cascades in transfected Chinese hamster ovary cells. *J Biol Chem* 267(4):2437-2442.

44. Takeda Y*, et al.* (1992) Ligand binding kinetics of substance P and neurokinin A receptors stably expressed in Chinese hamster ovary cells and evidence for differential stimulation of inositol 1,4,5-trisphosphate and cyclic AMP second messenger responses. *J Neurochem* 59(2):740-745.

45. Garcia M, Sakamoto K, Shigekawa M, Nakanishi S, & Ito S (1994) Multiple mechanisms of arachidonic acid release in Chinese hamster ovary cells transfected with cDNA of substance P receptor. *Biochem Pharmacol* 48(9):1735-1741.

46. Mitsuhashi M*, et al.* (1992) Multiple intracellular signaling pathways of the neuropeptide substance P receptor. *Journal of neuroscience research* 32(3):437-443.

47. Seabrook GR & Fong TM (1993) Thapsigargin blocks the mobilisation of intracellular calcium caused by activation of human NK1 (long) receptors expressed in Chinese hamster ovary cells. *Neurosci Lett* 152(1-2):9-12.

48. Grady EF*, et al.* (1995) Delineation of the endocytic pathway of substance P and its seven-transmembrane domain NK1 receptor. *Molecular biology of the cell* 6(5):509-524.

49. Segawa T, Nakata Y, Yajima H, & Kitagawa K (1977) Further observation on the lack of active uptake system for substance P in the central nervous system. *Jpn J Pharmacol* 27(4):573-580.

50. Heymann E & Mentlein R (1978) Liver dipeptidyl aminopeptidase IV hydrolyzes substance P. *FEBS Lett* 91(2):360-364.

51. Blumberg S, Teichberg VI, Charli JL, Hersh LB, & McKelvy JF (1980) Cleavage of substance P to an N-terminal tetrapeptide and a C-terminal

heptapeptide by a post-proline Cleaving enzyme from bovine brain. *Brain Res* 192(2):477-486.

52. Nyberg F, Le Greves P, Sundqvist C, & Terenius L (1984) Characterization of substance P(1-7) and (1-8) generating enzyme in human cerebrospinal fluid. *Biochem Biophys Res Commun* 125(1):244-250.

53. Nolte WM, Tagore DM, Lane WS, & Saghatelian A (2009) Peptidomics of prolyl endopeptidase in the central nervous system. *Biochemistry* 48(50):11971-11981.

54. Tenorio-Laranga J, Valero ML, Mannisto PT, Sanchez del Pino M, & Garcia-Horsman JA (2009) Combination of snap freezing, differential pH two-dimensional reverse-phase high-performance liquid chromatography, and iTRAQ technology for the peptidomic analysis of the effect of prolyl oligopeptidase inhibition in the rat brain. *Anal Biochem* 393(1):80-87.

55. Defendini R & Zimmerman EA (1993) Angiotensin converting enzyme in the brain. *J Neurochem* 60(2):787-789.

56. Yokosawa H, Endo S, Ogura Y, & Ishii S (1983) A new feature of angiotensin-converting enzyme in the brain: hydrolysis of substance P. *Biochem Biophys Res Commun* 116(2):735-742.

57. Defendini R, Zimmerman EA, Weare JA, Alhenc-Gelas F, & Erdos EG (1983) Angiotensin-converting enzyme in epithelial and neuroepithelial cells. *Neuroendocrinology* 37(1):32-40.

58. Defendini R, Zimmerman EA, Weare JA, Alhenc-Gelas F, & Edros EG (1982) Hydrolysis of enkephalins by human converting enzyme and localization of the enzyme in neuronal components of the brain. *Adv Biochem Psychopharmacol* 33:271-280.

59. Matsas R, Rattray M, Kenny AJ, & Turner AJ (1985) The metabolism of neuropeptides. Endopeptidase-24.11 in human synaptic membrane preparations hydrolyses substance P. *Biochem J* 228(2):487-492.

60. Kato T*, et al.* (1978) Successive cleavage of N-terminal Arg1--Pro2 and Lys3-Pro4 from substance P but no release of Arg1-Pro2 from bradykinin, by X-Pro dipeptidyl-aminopeptidase. *Biochim Biophys Acta* 525(2):417-422.

61. Horsthemke B*, et al.* (1984) Subcellular distribution of particle-bound neutral peptidases capable of hydrolyzing gonadoliberin, thyroliberin, enkephalin and substance P. *Eur J Biochem* 139(2):315-320.

62. Couture R & Regoli D (1981) Inactivation of substance P and its C-terminal fragments in rat plasma and its inhibition by Captopril. *Can J Physiol Pharmacol* 59(6):621-625.

63. Mauborgne A*, et al.* (1987) Enkephalinase is involved in the degradation of endogenous substance P released from slices of rat substantia nigra. *J Pharmacol Exp Ther* 243(2):674-680.

64. Mauborgne A, Bourgoin S, Benoliel JJ, Hamon M, & Cesselin F (1991) Is substance P released from slices of the rat spinal cord inactivated by peptidase(s) distinct from both 'enkephalinase' and 'angiotensin-converting enzyme'? *Neurosci Lett* 123(2):221-225.

65. Skidgel RA & Erdos EG (2004) Angiotensin converting enzyme (ACE) and neprilysin hydrolyze neuropeptides: a brief history, the beginning and follow-ups to early studies. *Peptides* 25(3):521-525.

66. Lee CM, Sandberg BE, Hanley MR, & Iversen LL (1981) Purification and characterisation of a membrane-bound substance-P-degrading enzyme from human brain. *Eur J Biochem* 114(2):315-327.

67. Probert L & Hanley MR (1987) The immunocytochemical localisation of 'substance-P-degrading enzyme' within the rat spinal cord. *Neurosci Lett* 78(2):132-137.

68. Karlsson K & Nyberg F (1998) Purification of substance P endopeptidase (SPE) activity in human spinal cord and subsequent comparative studies with SPE in cerebrospinal fluid and with chymotrypsin. *J Mol Recognit* 11(1-6):266-269.

69. Karlsson K, Eriksson U, Andren P, & Nyberg F (1997) Purification and characterization of substance P endopeptidase activities in the rat spinal cord. *Prep Biochem Biotechnol* 27(1):59-78.

70. Karlsson K & Nyberg F (2000) Purification of substance P endopeptidase activity in the rat ventral tegemental area with the Akta-Purifier chromatographic system. *J Chromatogr A* 893(1):107-113.

71. Matsas R, Kenny AJ, & Turner AJ (1984) The metabolism of neuropeptides. The hydrolysis of peptides, including enkephalins, tachykinins and their analogues, by endopeptidase-24.11. *Biochem J* 223(2):433-440.

72. Edwardson JA & McDermott JR (1985) Metabolism of neuropeptides at brain and pituitary sites. *Biochem Soc Trans* 13(1):50-53.

73. Sakurada T, Le Greves P, Stewart J, & Terenius L (1985) Measurement of substance P metabolites in rat CNS. *J Neurochem* 44(3):718-722.

74. Eriksson U, Andren PE, Caprioli RM, & Nyberg F (1996) Reversed-phase high-performance liquid chromatography combined with tandem mass spectrometry in studies of a substance P-converting enzyme from human cerebrospinal fluid. *J Chromatogr A* 743(1):213-220.

75. Kostel KL & Lunte SM (1997) Evaluation of capillary electrophoresis with post-column derivatization and laser-induced fluorescence detection for

the determination of substance P and its metabolites. *J Chromatogr B Biomed Sci Appl* 695(1):27-38.

76. Freed AL, Cooper JD, Davies MI, & Lunte SM (2001) Investigation of the metabolism of substance P in rat striatum by microdialysis sampling and capillary electrophoresis with laser-induced fluorescence detection. *J Neurosci Methods* 109(1):23-29.

77. Stewart JM*, et al.* (1982) A fragment of substance P with specific central activity: SP(1-7). *Peptides* 3(5):851-857.

78. Hall ME & Stewart JM (1992) The substance P fragment SP(1-7) stimulates motor behavior and nigral dopamine release. *Pharmacol Biochem Behav* 41(1):75-78.

79. Velazquez RA, Sun X, Kurtz HJ, & Larson AA (1993) Possible role of the N-terminus of substance P in kainic acid-induced toxicity in rats. *Brain Res* 624(1-2):109-114.

80. Hall ME, Miley F, & Stewart JM (1989) The role of enzymatic processing in the biological actions of substance P. *Peptides* 10(4):895-901.

81. Growcott JW & Tarpey AV (1982) Effects of substance P-(1-9) nonapeptide amide on inactivation of substance P in vitro. *Eur J Pharmacol* 84(1-2):107-109.

82. Khan S, Grogan E, Whelpton R, & Michael-Titus AT (1996) N- and C-terminal substance P fragments modulate striatal dopamine outflow through a cholinergic link mediated by muscarinic receptors. *Neuroscience* 73(4):919-927.

83. Khan S, Brooks N, Whelpton R, & Michael-Titus AT (1995) Substance P-(1-7) and substance P-(5-11) locally modulate dopamine release in rat striatum. *Eur J Pharmacol* 282(1-3):229-233.

84. Severini C, Improta G, Falconieri-Erspamer G, Salvadori S, & Erspamer V (2002) The tachykinin peptide family. *Pharmacol Rev* 54(2):285-322.

85. Deacon CF (2004) Circulation and degradation of GIP and GLP-1. *Horm Metab Res* 36(11-12):761-765.

86. Deacon CF, Johnsen AH, & Holst JJ (1995) Degradation of glucagon-like peptide-1 by human plasma in vitro yields an N-terminally truncated peptide that is a major endogenous metabolite in vivo. *J Clin Endocrinol Metab* 80(3):952-957.

87. Rissler K (1995) Sample preparation, high-performance liquid chromatographic separation and determination of substance P-related peptides. *J Chromatogr B Biomed Appl* 665(2):233-270.

88. Jankowski V*, et al.* (2007) Mass-spectrometric identification of a novel angiotensin peptide in human plasma. *Arterioscler Thromb Vasc Biol* 27(2):297-302.

89. Schulz-Knappe P*, et al.* (2001) Peptidomics: the comprehensive analysis of peptides in complex biological mixtures. *Comb Chem High Throughput Screen* 4(2):207-217.

90. Skold K*, et al.* (2002) A neuroproteomic approach to targeting neuropeptides in the brain. *Proteomics* 2(4):447-454.

91. Tagore DM*, et al.* (2009) Peptidase substrates via global peptide profiling. *Nat Chem Biol* 5(1):23-25.

92. Tinoco AD, Tagore DM, & Saghatelian A (2010) Expanding the dipeptidyl peptidase 4-regulated peptidome via an optimized peptidomics platform. *J Am Chem Soc* 132(11):3819-3830.

93. Tinoco AD*, et al.* (2011) A peptidomics strategy to elucidate the proteolytic pathways that inactivate peptide hormones. *Biochemistry* 50(12):2213-2222.

Chapter 3

Elucidating Substance P Regulation in the Spinal Cord

Dr. Arthur Tinoco and I performed the peptidomics experiments, the cross-linking experiments and the inhibitor injection experiments. Dr. Arthur Tinoco performed the inhibitor screens. I synthesized and purified the peptides and performed the in vitro degradation assays.

3.1 Introduction

Substance P is an 11-amino acid bioactive peptide that functions as a neurotransmitter in the mammalian nervous system (1). It is one of the most widely studied bioactive peptides and has been shown to participate in a variety of biological processes, including pain transmission (2-4), inflammation (5, 6) and depression (7-9). However, despite the effort that has gone into uncovering the biology of substance P, neither the molecular pathways that control its activity nor the enzymes that drive them have been definitively identified[2].

Historically, investigations of bioactive peptide metabolism have relied on *in vitro* experiments to glean insights into endogenous degradation pathways. While these experiments can be very informative, they are also difficult to interpret because the cellular environment in which the endogenous peptide-peptidase interactions occur may not be perfectly reconstituted in the test tube. Our lab has developed a general approach to elucidating bioactive peptide metabolism that we believe circumvents this challenge. By incorporating LC-MS/MS-based peptidomics experiments into our core workflow, we achieve a view directly into the *in vivo* metabolism of the peptide and can thereby identify physiologically relevant metabolites. Subsequent *in vitro* studies can then be used to tease apart the molecular pathways that generate these products.

---

[2] Inhibition of prolyl endopeptidase (PREP), which acts on the N-terminus of substance P to produce SP(3-11) and SP(5-11) , slightly increases substance P levels in the brain (11, 12), which suggests that the enzyme may degrade substance P *in vivo*. However, because SP(3-11) and SP(5-11) are both very good agonists for the primary substance P receptor neurokinin 1, it is unlikely that PREP is the primary inactivator of substance P *in vivo*.

Our laboratory recently demonstrated the effectiveness of this approach by characterizing the intestinal metabolism of peptide histidine isolucine (PHI), a hormone that promotes glucose-stimulated insulin secretion (13). Here I describe a study in which we brought such an approach to bear on substance P metabolism in the spinal cord. Specifically, we harnessed our laboratory's *in vivo* peptidomics expertise to identify two physiological relevant metabolites of substance P in the spinal cord: the N-termial fragments SP(1-9) and SP(1-7). Focusing our efforts on the SP(1-9)-producing pathway, we then utilized *in vitro* biochemical assays to identify a GM6001-sensitive activity that generate SP(1-9) from substance P and a GM6001-sensitive activity that constitutes a majority of the total substance P-degrading activity in spinal cord. Finally, we determine that GM6001 treatment causes a nearly three-fold increase in endogenous substance P levels in the spinal cord. This is the largest change in substance P levels ever induced by a genetic or pharmacological strategy and indicates that GM6001 blocks a pathway that controls the endogenous levels of substance P in the spinal cord.

3.2 Quantitative *in vivo* peptidomics analysis of mouse spinal cord to identify physiological metabolites of substance P

Our first goal was to identify physiologically relevant substance P degradation fragments. To achieve this, we extracted the mouse spinal cord peptidome using an LC-MS/MS-based peptidomics platform that our laboratory had previously

developed (11, 14, 15). We detected only full-length substance P and the N-terminal fragments SP(1-9) and SP(1-7) (Figure 3.1). Subsequent isotope dilution mass spectrometry (IDMS) experiments determined that the *in vivo* concentration of substance P, SP(1-9) and SP(1-7) were 105.86 ± 8.53 pmol/g, 2.07 ± 0.48 pmol/g and 1.63 ± 0.50 pmol/g, respectively (Table 3.1). The quantity of the full-length peptide is comparable to the amount found in rat spinal cord (16). While these results do not rule out the possibility that other substance P fragments are present *in vivo*, they do indicate that SP(1-9) and SP(1-7) are endogenous substance P metabolites.

There are two simple pathway models that could explain our *in vivo* observations. In one model, SP(1-9) and SP(1-7) are produced by independent pathways (Figure 3.2A). In the other, SP(1-9) and SP(1-7) are produced sequentially in a single pathway (Figure 3.2B). The metabolic step from substance P to SP(1-9) takes place in both models and we know from previous studies that this step would amount to an inactivation substance P, as the SP(1-9) fragment does not bind to the substance P receptor NK_1 (17, 18). We therefore decided to focus our efforts on understanding the substance P to SP(1-9) pathway.

Figure 3.1 Quantitative *in vivo* peptidomics analysis of mouse spinal cord. Substance P (A) and all detectable metabolites (B and C) were quantified by isotope dilution-mass spectrometry (IDMS). We synthesized deuterated (heavy-labeled) variants of the peptides in question and added known quantities of same to mouse spinal cord prior to our peptidomics analysis. The peptides are then quantified by LC-MS; the deuterated SEP and endogenous SEP can be distinguished by differences in their masses.

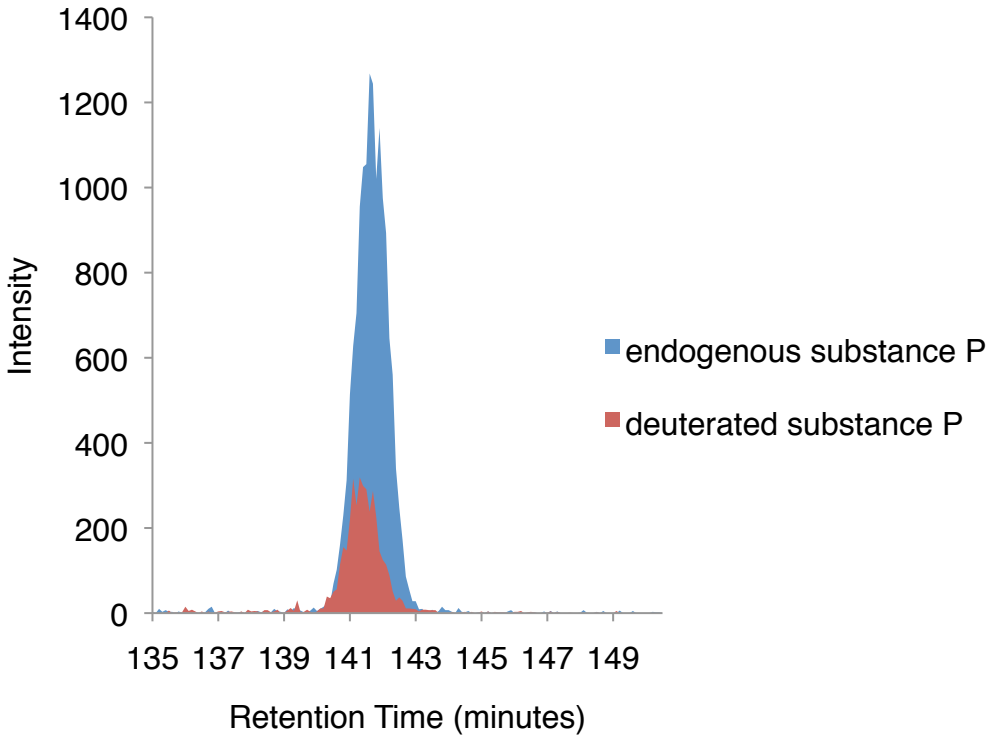


Figure 3.1 (Continued)

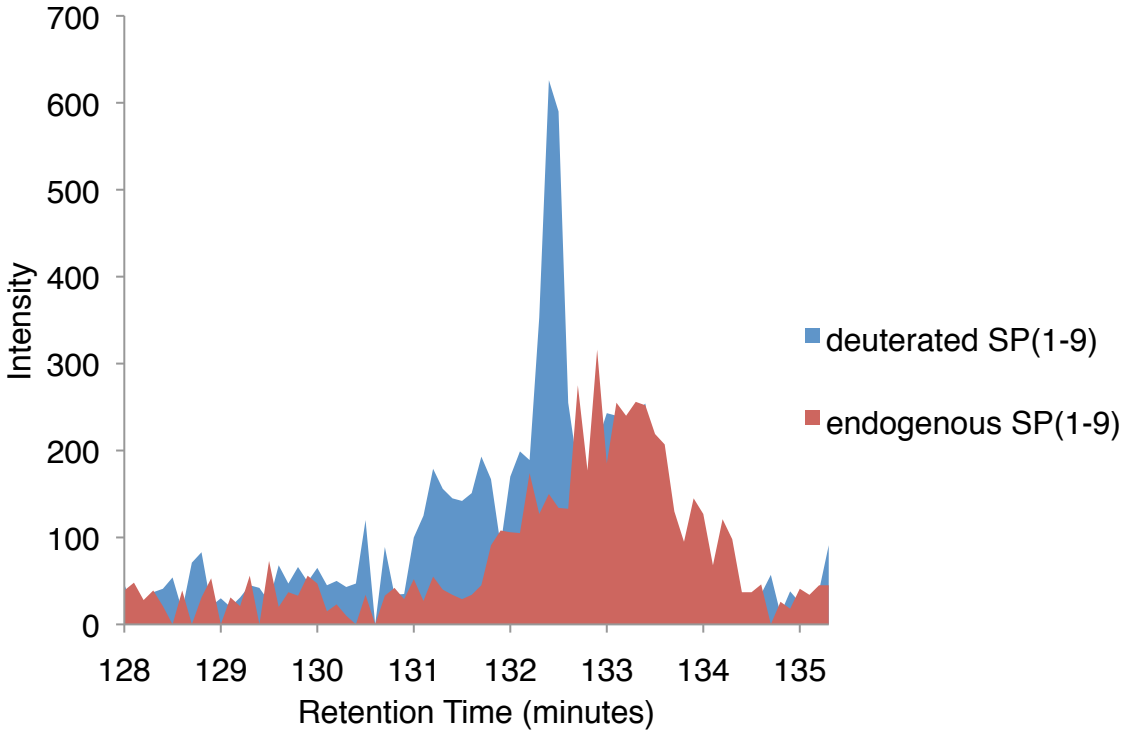


Figure 3.1 (Continued)

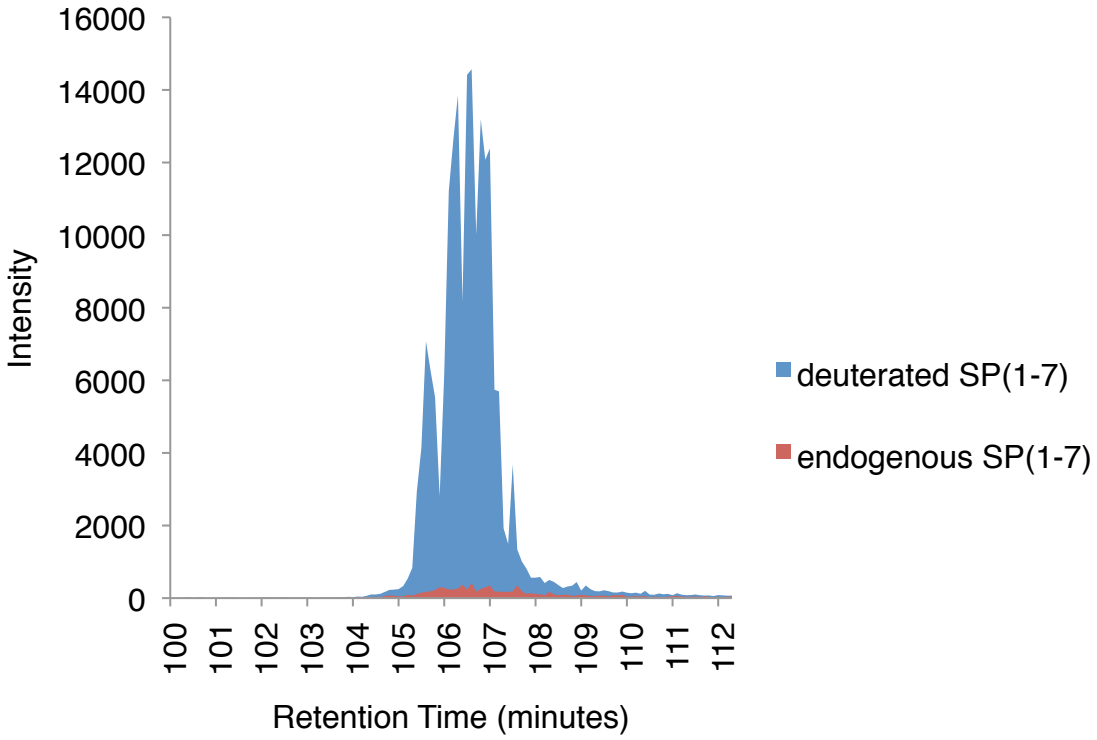


Figure 3.1 (Continued)

Table 3.2 Concentrations of substance P, SP(1-9) and SP(1-7) in mouse spinal cord.

Peptide	Concentration (pmol/g)
Substance P	105.86 ± 8.53
SP(1-9)	2.07 ± 0.48
SP(1-7)	1.63 ± 0.50

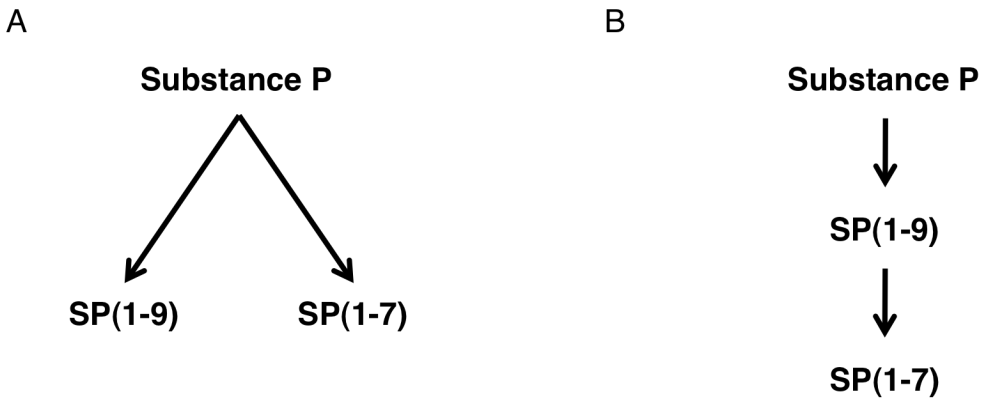


Figure 3.2 Possible substance P degradation pathways. (A) SP(1-9) and SP(1-7) are produced by independent pathways. (B) SP(1-9) and SP(1-7) are produced sequentially in a single pathway. Intermediates not shown here could exist.

3.3 *In vitro* degradation assay with mouse spinal cord lysate

We next performed an *in vitro* biochemical assay to determine whether the enzymes responsible for producing these fragments are present in the spinal cord (it is otherwise possible that the full length peptide is processed elsewhere and the metabolites then transported to the spinal cord). In this experiment, we incubated the soluble and insoluble fractions of mouse spinal cord lysate with synthetic substance P peptide and then analyzed the quenched reactions by matrix assisted-laser desorption ionization (MALDI)-time of flight (TOF)-MS. The sensitivity of this instrument is such that we will only observe peptide fragments produced from degradation of the exogenous peptide, which is added at a quantity well above the detection limit, and not fragments produced by endogenous metabolism. We observed that the insoluble fraction of the lysate produced SP(1-10), SP(1-9), SP(1-8), and SP(1-7) (Figure 3.3A) whereas the soluble fraction did not produce detectable amounts of any N-terminal fragment peptide (Figure 3.3B). This indicates that there is enzyme activity present in the spinal cord capable of producing the fragments we detect *in vivo* and that a membrane or membrane-bound enzyme is responsible for the activity, which is expected given that substance P is degraded in the extracellular space.

Figure 3.3 *In vitro* experiment with either the (A) insoluble or (B) soluble fraction of mouse spinal cord lysate. Lysate (1 mg/ml total protein) was incubated with substance P (100 uM) for 1 hour at 37 C. (A) Substance P degrading activity in the insoluble fraction is competent to produce SP(1-10), SP(1-9), SP(1-8), and SP(1-7). (B) No N-terminal substance P fragments were detected in the soluble experiment.

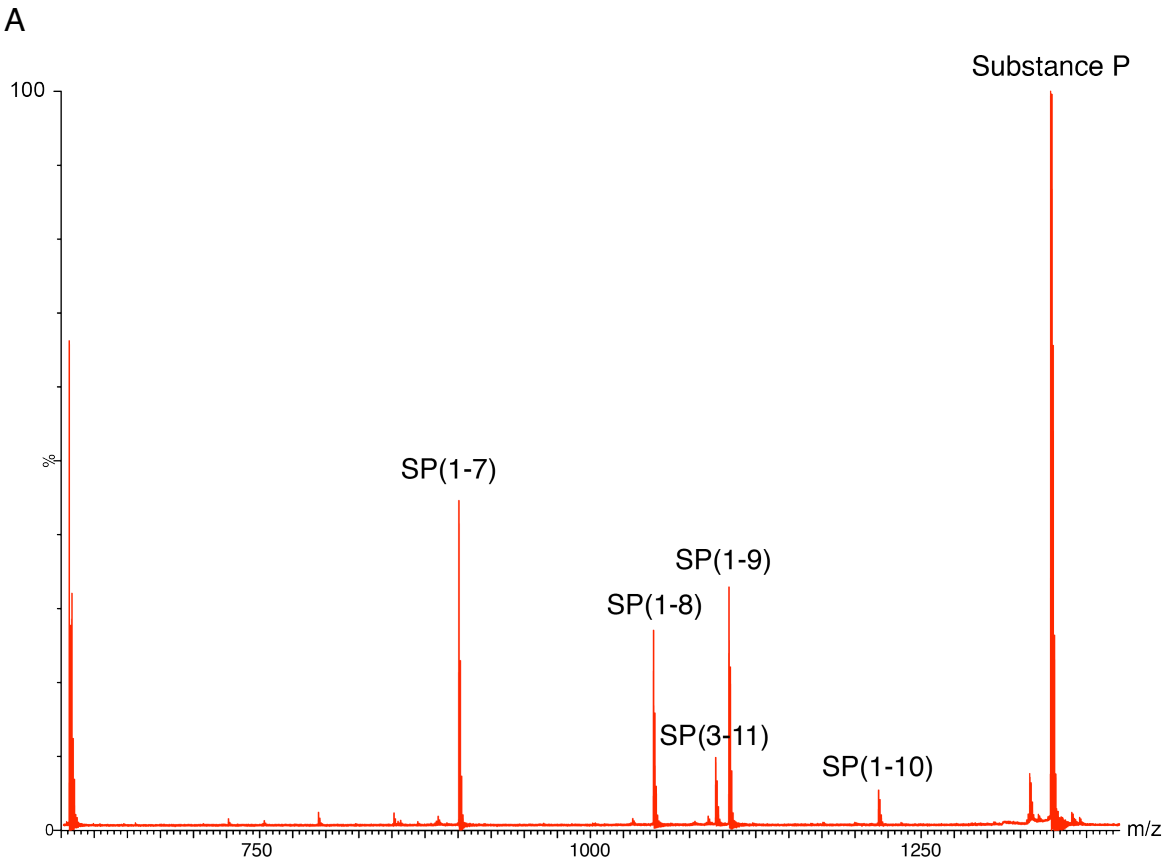


Figure 3.3 (Continued)

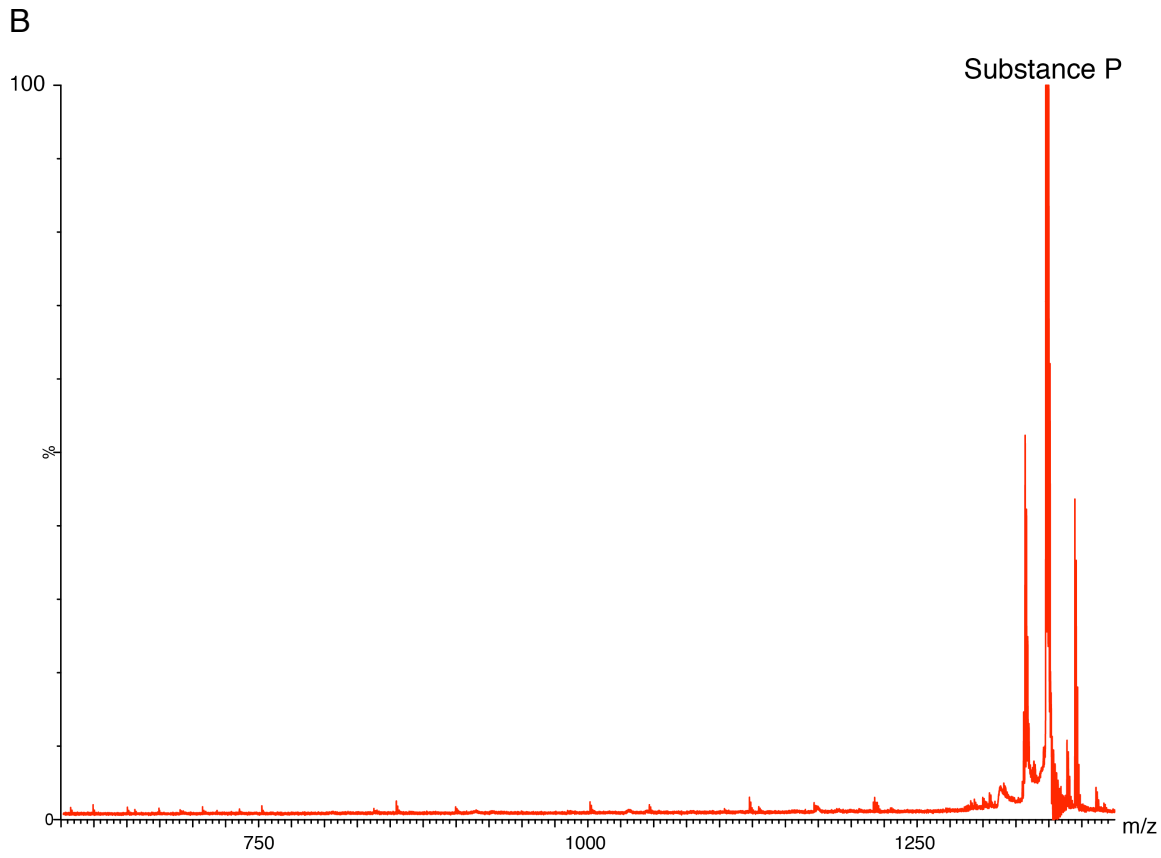


Figure 3.3 (Continued)

3.4 Assembling a candidate enzyme list from the MEROPS database

The MEROPS database is an authoritative catalogue of all known peptidases and includes detailed information on each peptidase's specificity (19). Because substance P is extremely stable in solution and is also very widely studied, it is one of the standard substrates used in specificity assays in studies aimed at biochemically characterizing a peptidase. Thus, it is likely that most if not all of the known peptidases that are capable of cleaving substance P are known to be capable of cleaving substance P. We therefore thought that a list of enzymes annotated in MEROPS to be capable of cleaving substance P at the Gly^{9}-Leu^{10} amide bond would be a good initial candidate list of enzymes that could be responsible for the endogenous SP(1-9) producing activity. If the enzyme responsible for the activity were already characterized, starting from this approach would decrease the time to discovery because it is easier to eliminate candidates using inhibitor-based approaches than to make positive identifications in an unbiased search for an enzyme. Moreover, given the numerous annotated peptidases that have been put forth as endogenous substance P degrading enzymes (20-28), determining that a non-annotated enzyme drove a substance P-degradation pathway would be a significant result in itself.

A search of the MEROPS database revealed that a total of nine mammalian peptidases are capable of cleaving substance P to produce SP(1-9) (Figure 3.4). We reduced this list to a candidate set of seven by eliminating enzymes that had extremely low abundance or else were not present in the

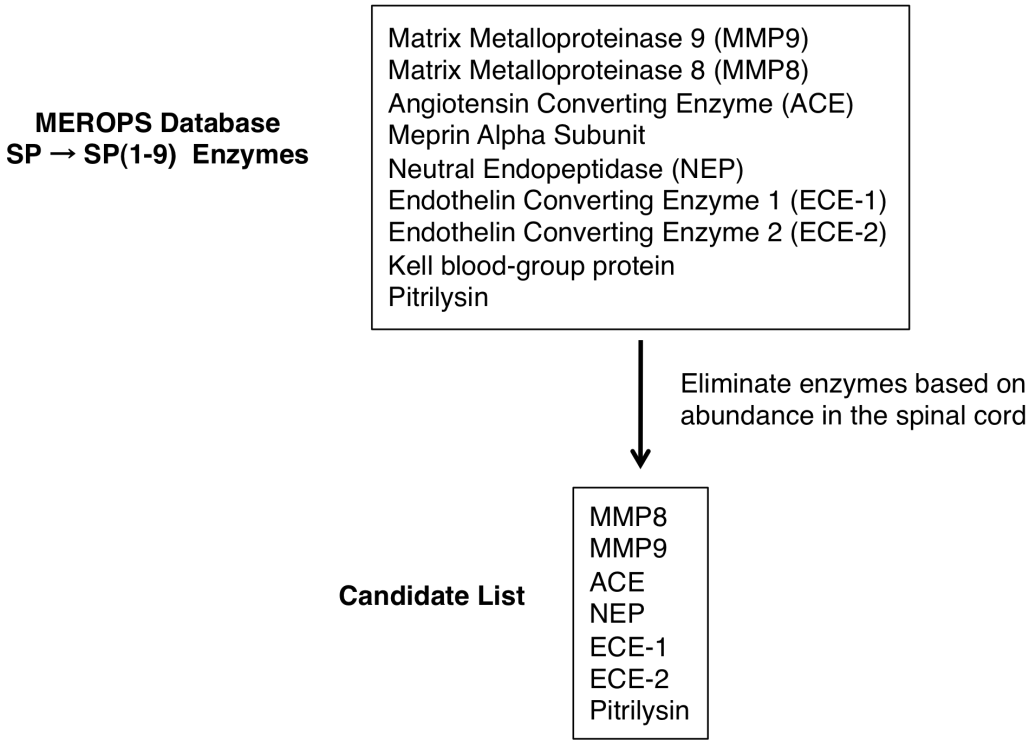


Figure 3.4 Assembling a candidate list SP(1-9)-producing enzymes using compendiums of peptidase information. A search of the MEROPS database revealed that a total of nine mammalian peptidases are capable of cleaving substance P to produce SP(1-9). We reduced this list to a candidate set of five by eliminating enzymes that had extremely low abundance or else were not present in the spinal cord as indicated in the Allen Mouse Spinal Cord Atlas (29) or the NEXTBIO Research Database (http://www.nextbio.com/b/nextbio.nb).

spinal cord (29). Of these seven enzymes, one (pitrilysin) was annotated as not being a membrane or membrane-bound protein. We decided to keep this enzyme in our candidate list despite the fact that the SP(1-9)-producing activity is present in the insoluble fraction of mouse spinal cord lysate because we thought it possible that pitrilysin could be membrane associated despite the annotation.

3.5 Class-specific protease inhibitor screening to identify candidate enzyme class

To determine the class of the peptidase responsible for the SP(1-9)-producing activity, we incubated the insoluble fraction of mouse spinal cord lysate (1 mg/ml total protein concentration) with synthetic substance P (100 uM) in the presence of a battery of class-specific protease inhibitors. We found that the activity was sensitive only to metallopeptidase inhibitors and in particular to the zinc chelator O-phenanthroline (Figure 3.5B). This finding is consistent with our candidate list containing the true SP(1-9)-producing enzyme, as every enzyme in the list is a metallopeptidase.

Interestingly, in addition to blocking the formation of SP(1-9), this inhibitor significantly reduces the degradation of substance P, which suggests it is blocking a major degradation pathway. We also observe that o-phenanthroline completely eliminates SP(1-7) formation (Figure 3.5A). However, we cannot be certain that this is due to a reduction in the levels of SP(1-9) since it is possible that GM6001 is inhibiting another enzyme that drives an alternative SP(1-7) pathway (i.e., not an SP(1-9) to SP(1-7) pathway). This is consistent with SP(1-9)

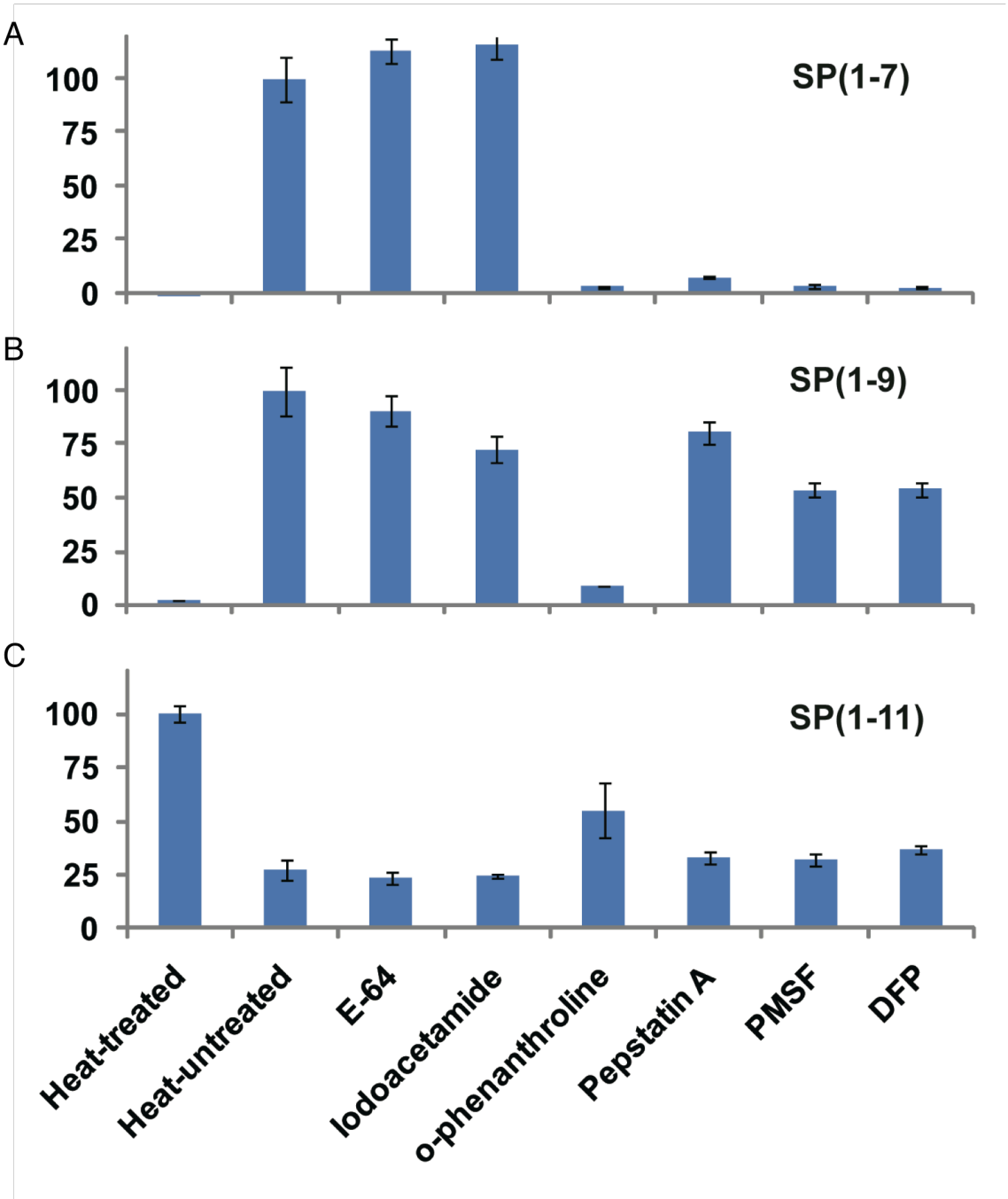


Figure 3.5 Sensitivity of substance P-degrading activity to class-specific peptidase inhibitors. Metallopeptidase inhibitor o-phenanthroline blocks the formation of (A) SP(1-7) and (B) SP(1-9) while significantly reducing substance P degradation (C). In the case of substance P degradation and SP(1-9) formation, o-phenanthroline has a much greater effect than any other inhibitor. In (A) and (B), the vertical axis is normalized to the heat untreated sample. In (C), the vertical axis is normalized relative to the heat-treated sample.

and SP(1-7) being the products of a single pathway. Pepstatin A, which is an aspartylpeptidase inhibitor, and PMSF and DMF, which are serine peptidase inhibitors, also abolish SP(1-7) formation, which suggest that there are multiple SP(1-7)-producing activities present in the spinal cord.

3.6 Enzyme-specific protease inhibitor screening to evaluate candidate enzymes

To determine whether any of the enzymes in our candidate list were responsible for the SP(1-9)-producing activity, we incubated the insoluble fraction of mouse spinal cord lysate (1 mg/ml total protein concentration) with synthetic substance P (100 uM) in the presence of a battery of class-specific protease inhibitors. For each candidate enzyme, we included in the survey at least one inhibitor that is capable of inhibiting that enzyme and not capable of inhibiting any other candidate enzyme (with exception of NEP and ECE-2; neither of these enzymes is sensitive to an inhibitor that does not inhibit the other enzyme so we used phosphoramidon, which is a potent inhibitor of both enzymes).

We found that GM6001, which inhibits enzymes in the neprilysin and MMP families, substantially prevents degradation of substance P, with peptide levels being over 2-fold higher in the treated sample than in the control. No other inhibitors substantially prevented substance P degradation (Figure 3.6C). Moreover, GM6001 completely blocks formation of SP(1-9), which reinforces the link between substance P levels and the SP(1-9) pathway (Figure 3.6 B).

Figure 3.6 Sensitivity of substance P-degrading activity to enzyme-specific peptidase inhibitors. Inhibitor-enzyme pairs: SM-19712 (ECE1), phosphoramidon (neprilysin; ECE2), MMP9 inhibitor, TIMP2 (MMP inhibitor), GM6001 (MMP and neprilysin broad inhibitor), chymostatin (pitrilysin), captopril (ACE), enalaprilat (ACE), actinonin (Meprin 1A), and vehicle (PBS with DMSO for 5% DMSO final concentration in reaction). MMP and neprilysin family inhibitor GM6001 blocks the formation of (A) SP(1-7) and (B) SP(1-9) while significantly reducing substance P degradation (C). In the case of substance P degradation and SP(1-9) formation, GM6001 has a much greater effect than any other inhibitor. In (A) and (B), the vertical axis is normalized to the heat untreated sample. In (C), the vertical axis is normalized relative to the heat-treated sample.

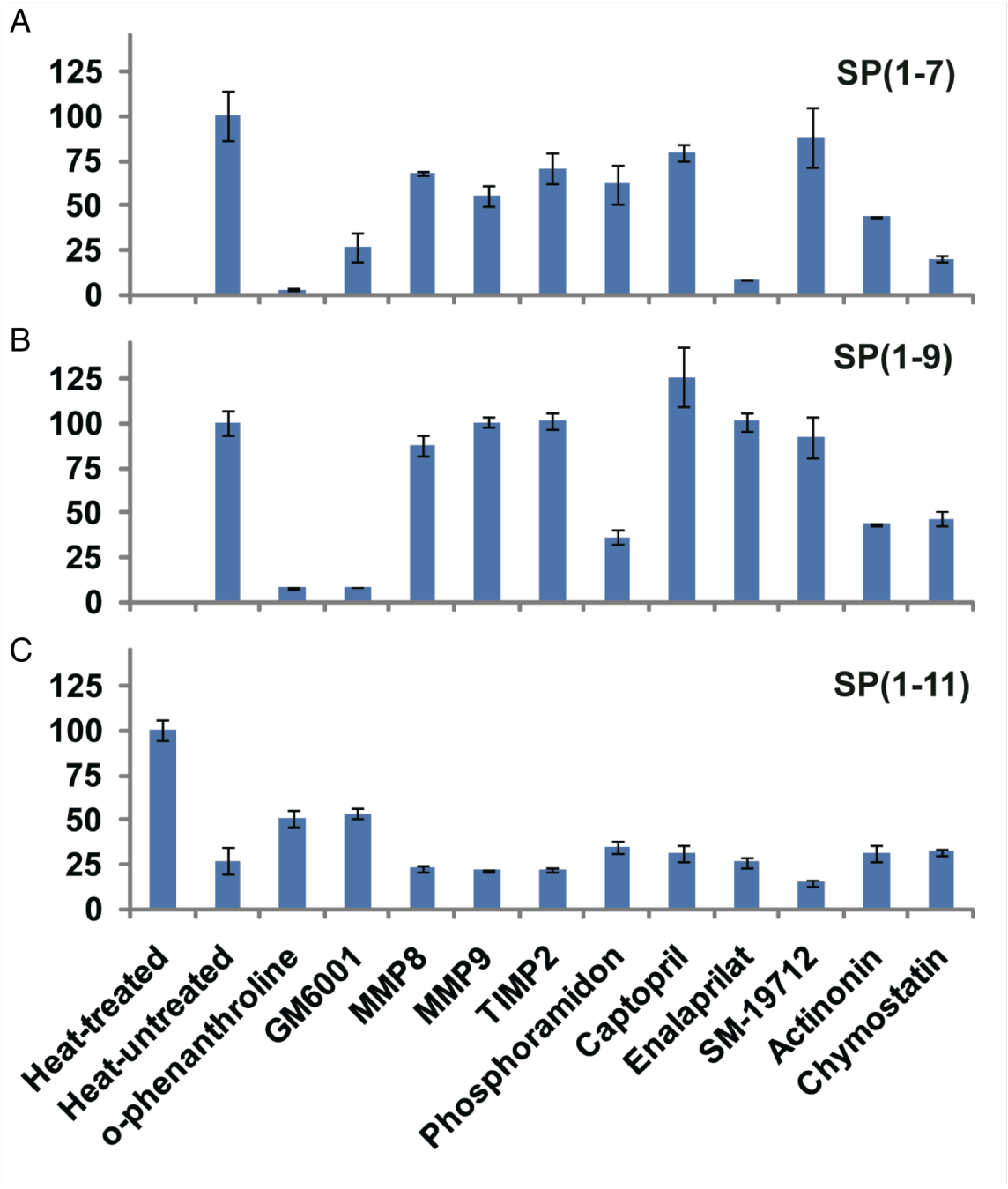


Figure 3.6 (Continued)

Phosphoramidon, which inhibits neprilysin and ECE-2, and chymostatin, which inhibits pitrilysin, also prevent SP(1-9) formation to an appreciable degree, though not nearly to the extent that GM6001 does (Figure 3.6B). This could be due to cross inhibition of a single enzyme by GM6001, phosphoramidon and chymostatin, where the latter two are weaker inhibitors for the enzyme. This possibility seems especially plausible in the case of GM6001 and phosphoramidon considering that both inhibitors target M13 family peptidases.

Alternatively, there could be more than one pathway producing SP(1-9) involving NEP, pitrilysin or both. In either case, however, the GM6001-inhibited enzyme appears to be responsible for the primary SP(1-9)-producing activity and has by far the most significant impact on overall substance P levels of the inhibitors tested. We therefore eliminate NEP, ECE-2 and pitrilysin as candidates for the primary SP(1-9)-producing enzyme. We also eliminate MMP8, MMP9, ECE-1 or ACE because no significant change in substance P levels or SP(1-9) levels was observed for the reactions in which inhibitors to these enzymes were used.

3.7 Wild type vs NEP$^{-/-}$ comparative study

Prompted by the inconclusive results of the inhibitor survey with regard to NEP, we wished to explore the possibility that NEP is involved in substance P metabolism, either by driving the SP(1-9) pathway or otherwise. We therefore compared the substance P-degrading properties of wild type mouse spinal cord tissue with that of lysate from the spinal cords of mice lacking the neprilysin gene

(NEP$^{-/-}$) using our *in vitro* assay. We found that there was no difference in overall subtstance P levels between the reactions and only a negligible difference in SP(1-9) levels, indicating that NEP is not responsible for the SP(1-9)-producing activity we had observed (Figure 3.7A).

We wished to confirm that NEP was not responsible for controlling substance P levels *in vivo* so we used IDMS to quantify the levels of substance P in wild type and NEP$^{-/-}$. We determined that substance P levels in the spinal cord are identical in wild type and NEP$^{-/-}$ mice (Figure 3.7B) and conclude that NEP is not involved in controlling substance P levels in the spinal cord.

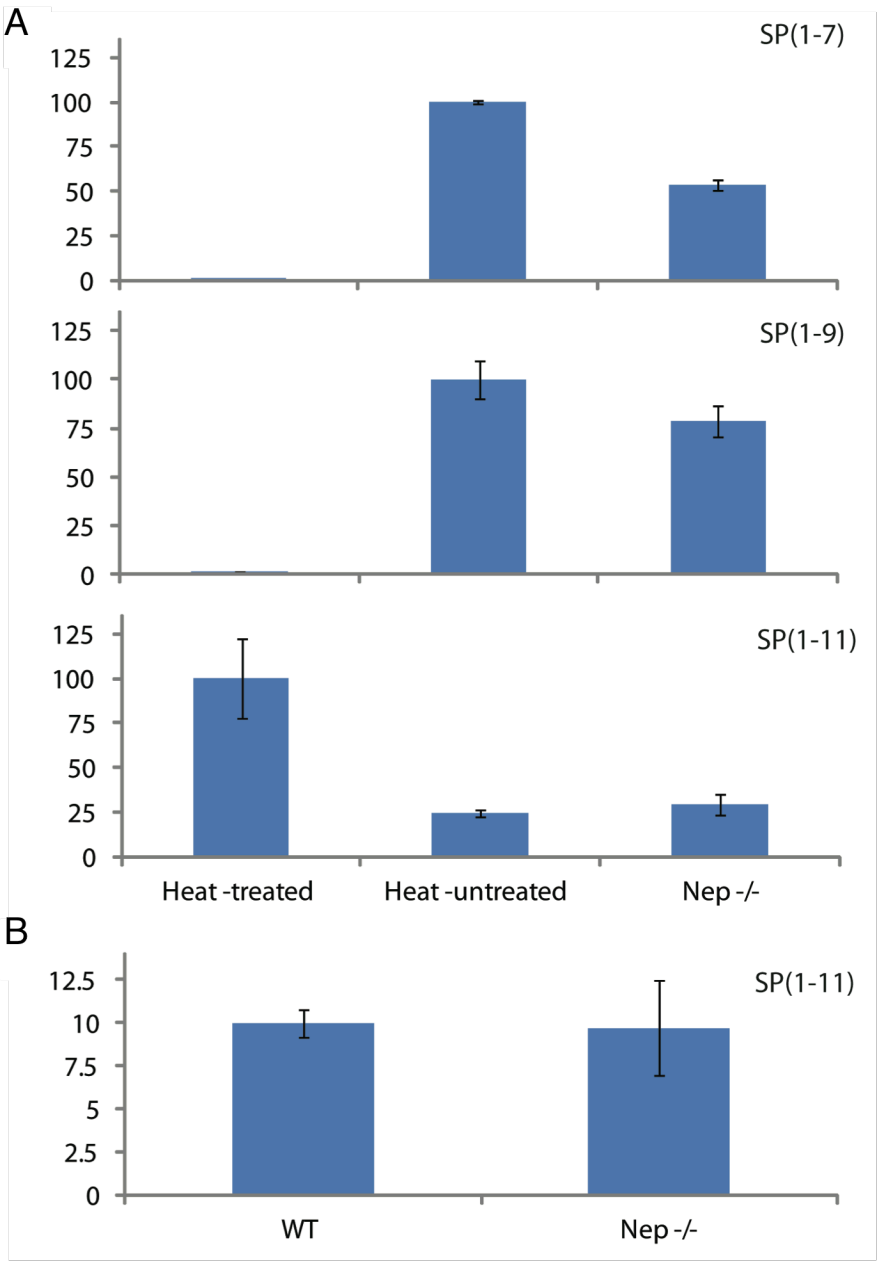


Figure 3.7 Comparison of substance P-degrading activity of wild type and NEP$^{-/-}$ mouse spinal cords: (A) *in vitro* and (B) *in vivo*. (A) There is a very small (~5%) difference in overall substance P-degrading activity between the wild type and NEP$^{-/-}$ spinal cord lysates (lower panel), indicating that NEP activity does not constitute a significant fraction of the total substance P-degrading activity in the spinal cord. SP(1-9) and SP(1-7) formation are reduced by approximately 20% and 40%, respectively, indicating that NEP contributes to the formation of these products in the *in vitro* assay. (B) Substance P levels in the spinal cord are identical in wild type and NEP$^{-/-}$ mice, indicating that NEP is not involved in controlling substance P levels in the spinal cord.

3.8 Cross-linking experiments using activity-based probes reveals that two enzymes may be responsible for the GM6001-sensitive activity

It is possible that more than one enzyme is responsible for the GM6001-sensitvie substance P-degrading activity that we observe in the spinal cord. To get a better sense of the number of enzymes that could be involved and get information about their respective size, we performed a competitive cross-linking experiment with a rhodamine-tagged hydroxamate benzophenone probe (HxBp-RH) that is derived from and has the same metal-binding moiety as GM6001 (30). In this experiment, spinal cord lysates were mixed with 100 nM of HxBP-Rh in the presence or absence of 100 μM GM6001 and then irradiated at 365 nm for 1 hour to induce cross-linking with the Bp structure of the label. In the samples in which it is present, GM6001 will outcompete HxBP-Rh at sites where it specifically binds and thus prevent fluorescence of enzymes that had been bound to HxBP-Rh in the GM6001-less sample. Thus, bands that fluoresce in the absence of GM6001 but disappear when it is added represent GM6001-binding proteins.

We observe that two proteins, one of molecular weight 100 kDa and the other of molecular weight 80 kDa, are bound by GM6001 (Figure 3.8). These enzymes could be responsible for the GM6001-inhibited substance P-degrading activity we observe. Of the remaining candidate enzymes, only ECE-2, which has a molecular weight of 100 kDa, could potentially be one of the proteins observed in the experiment. However, considering that the potent ECE-2 inhibitor

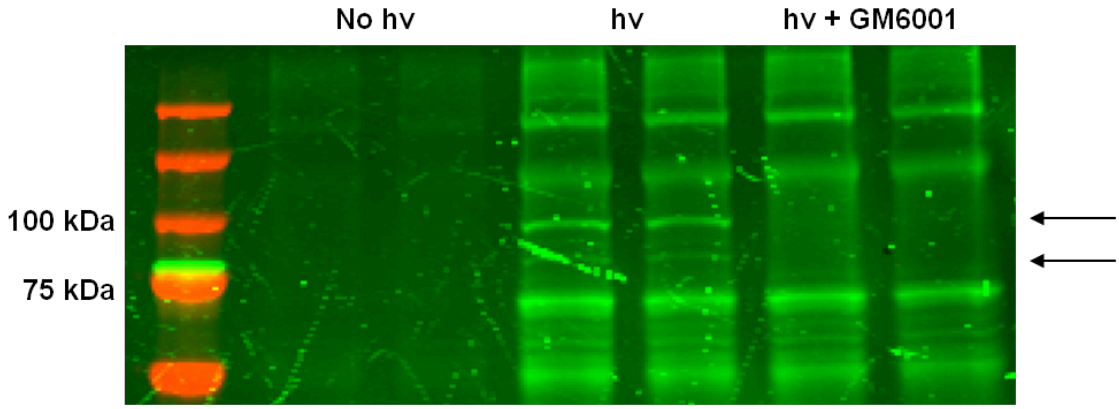


Figure 3.8 Determining the number and size of proteins responsible for the GM6001-sensitive substance P-degrading activity. We performed a competitive cross-linking experiment with a rhodamine-tagged hydroxamate benzophenone probe (HxBp-RH) that is derived from and has a nearly identical reactivity profile to GM6001 (30). Spinal cord lysates were mixed with 100 nM of HxBP-Rh in the presence (first two lanes from right) or absence (third and fourth lanes from right) of 100 μM GM6001 and then irradiated at 365 nm for 1 hour to induce fluorescence. A control was performed in the absence of GM6001 and no irradiation (first two lanes from left after ladder). In the samples in which it is present, GM6001 will outcompete HxBP-Rh and thus prevent fluorescence of enzymes that had been bound to HxBP-Rh in the GM6001-less sample. Thus, bands that fluoresce in the absence of GM6001 but disappear when it is added represent GM6001-binding proteins. We observe that two proteins, one of molecular weight 100 kDa and the other of molecular weight 80 kDa, are bound by GM6001 (Figure 3.8) and thus could be responsible for the GM6001-sensitive activity.

phosphoramidon does not significantly reduce the substance P-degrading activity we observe, if the 100 kDa band does correspond to ECE-2, then it is likely that only one enzyme – the 80 kDa protein - is responsible for the substance P-degrading activity we observe. As all of the enzymes known to cleave substance P to produce SP(1-9) have been ruled out as candidates, it is likely that this 80 kDa protein is a novel substance P-degrading enzyme.

3.9 Treatment with GM6001 significantly alters endogenous substance P levels

To determine whether the GM6001-sensitive substance P-degrading activity we were observing in our *in vitro* experiments was involved in controlling substance P levels *in vivo*, we injected 3-4 month old wild type mice with GM6001 to a concentration of 100mg of inhibitor to 1 kg of mouse (n=4). After 3 hours, the mice were sacrificed and the concentration of substance P in their spinal cords was determined by isotope dilution mass spectrometry. The substance P levels were nearly 2.73 times higher in the treated mice (Figure 3.9), indicating that GM6001 blocks a pathway that controls substance P levels in the spinal cord. This is the largest change in endogenous substance P levels that has even been induced, pharmacologically or genetically.

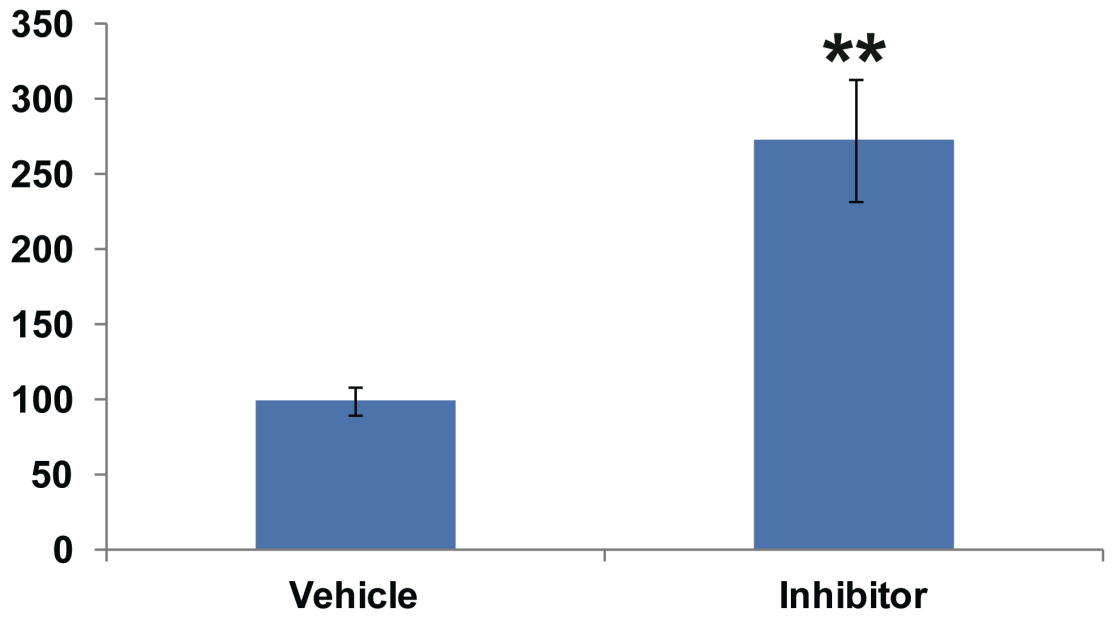


Figure 3.9 Measuring the impact of GM6001 treatment on substance P levels in the spinal cord. GM6001 was injected into 3-4 month old wild type mice to a concentration of 100mg of inhibitor to 1 kg of mouse (n=4). After 3 hours, the mice were sacrificed and the concentration us substance P in their spinal cords was determined by isotope dilution mass spectrometry. The substance P levels were nearly 2.73 times higher in treated mice than in untreated mice, which indicates that GM6001 blocks a pathway that controls substance P levels in the spinal cord. ("**" indicates a p-value of less than 0.01.)

3.10 Conclusion

In this study, we used an LC-MS/MS peptidomics technique to identify two physiologically relevant metabolites of substance P in the spinal cord: the N-termial fragments SP(1-9) and SP(1-7). Focusing our efforts on the pathway that produces SP(1-9), we then used *in vitro* biochemical assays to identify one or more GM6001-sensitive activities that generate SP(1-9) from substance P and, further, constitute a significant fraction of the total substance P-degrading activity in spinal cord lysate. With a competitive cross-linking strategy featuring a GM6001-like fluorescent probe, we find that two proteins, one of which is 100 kDa and other of which is 80 kDa, may be involved in producing this activity. We also determine that none of the enzymes currently known to cleave substance P to produce SP(1-9) are involved.

Significantly, we also find that GM6001 treatment causes a nearly three-fold increase in endogenous substance P levels in the spinal cord. This is the largest change in substance P levels ever induced by a genetic or pharmacological strategy and indicates that GM6001 blocks a pathway that controls the endogenous levels of substance P in the spinal cord. GM6001 can therefore be used as an experimental tool to modulate substance P levels in an animal model, which could be useful in studies of conditions in which substance P levels are speculated to be raised, such as chronic pain and inflammation.

3.11 Materials and methods

Compounds

Mouse substance P was purchased from Anaspec, Inc. A protease inhibitor panel was obtained from Sigma Aldrich Inc.

Peptide synthesis

Heavy-labeled SP(1-7) (Pro containing five ^{13}C and one ^{15}N), SP(1-9) (Phe containing eight ^{2}H), and substance P (Leu containing ten ^{2}H) were synthesized manually using FMOC chemistry for solid-phase peptide synthesis. Crude peptides were purified by RP-HPLC (Shimadzu) using a C18 column (150 × 20 mm, 10 µm particle size, Higgins Analytical). The HPLC gradient varied depending on the peptide (Mobile Phase A: 99% H_2O, 1% Acetonitrile, 0.1% TFA; Mobile Phase B: 90% Acetonitrile, 10% H_2O, 0.07% TFA). HPLC fractions were analyzed for purity by MALDI-TOF (Waters) using α-cyano-4-hydroxycinnamic acid as the matrix. Pure fractions were combined and lyophilized. Concentrations of the purified peptides were determined by UV-vis using the extinction coefficient for Phe.

Animal studies

Wild type (C57BL/6) mice used in this study were either purchased (Jackson Labs) or taken from a breeding colony. *Nep*$^{-/-}$ mice were obtained from Craig

Gerard at Children's Hospital (Boston, MA) and were on a C57BL/6 background. Mice in these studies were not littermates from het x het crosses, but were obtained from separate colonies of *Nep*$^{-/-}$ and WT mice. All mice used in this study ranged from 3 to 6 months old. Animals were kept on a 12-h light, 12-h dark schedule and fed *ad libitum.* For spinal cord tissue collection, animals were euthanized with CO_2, their tissue dissected, flash frozen with liquid N_2, and stored at -80 °C. All animal care and use procedures were in strict accordance with the standing committee on the use of animals in research and teaching at Harvard University and the National Institute of Health guidelines for the humane treatment of laboratory animals.

Isolation of physiological peptides from tissue

Tissue peptide isolation and fractionation were previously described (15, 31). Briefly, frozen spinal cords were placed in 500 μL of water and boiled for 10 minutes to inactivate any residual proteolytic activity prior to tissue homogenization. The aqueous fraction was separated and saved, and the tissue was dounce-homogenized in ice-cold 0.25% aqueous acetic acid. The aqueous fraction and the homogenate were combined and centrifuged at 20,000 x g for 20 min at 4 °C. The supernatant was then sent through a 10 kDa molecular weight cut-off filter (VWR Modified PES) to enrich the peptide pool and then a C18 Sep Pak cartridge (HLB 1cc; 30 mg, Oasis) to desalt the sample. The peptides were

then eluted with 1 mL of 70:30 H_2O/ACN and concentrated under vacuum using a speed vac prior to fractionation by strong cation exchange (SCX).

SCX was performed using a PolySULFOETHYL A^{TM} column (200 x 2.1mm, 5 μm, 300 Å; PolyLC INC.) connected to an Agilent Technologies 1200 series LC. All runs were operated at 0.3 mL/min. The SCX buffers (prepared with MS quality water) consisted of: A) 7 mM KH_2PO_4, pH 2.6, 25% ACN (vol/vol); B) 40 mM KCl, 7 mM KH_2PO_4, pH 2.6, 25% ACN (vol/vol); C) 100 mM KCl, 7 mM KH_2PO_4, pH 2.6, 25% ACN (vol/vol); D) 600 mM KCl, 7 mM KH_2PO_4, pH 2.6, 25% ACN (vol/vol). Prior to the SCX runs, all samples (N = 4) were dissolved in 900 μL buffer A (1 mL sample loop). A step-gradient was applied that included 60 min with Buffer A, 60 min with Buffer B, 60 min with Buffer C, and 60 min with Buffer D, with 1 min transitions between the different buffer conditions. Fractions were collected separately for each of the different buffer conditions (e.g., a buffer A fraction, a buffer B fraction, and so on). Fraction C was isolated because substance P and all C-terminal peptide products are expected to be +3 charged at pH 2.6. This fraction was applied to a C18 Sep Pak cartridge, washed with water to desalt the samples, and then eluted with 1 mL of 70:30 H_2O/ACN and concentrated using a speed vac. It is important to note that the Met on substance P becomes nearly 100% oxidized following SCX. The peptide samples were dissolved in 0.1% aqueous formic acid (50 mg tissue/20 μL), normalized according to the original tissue weight, prior to LC-MS analysis.

LC-MS/MS experiments to detect SP c-terminal peptide fragments

Fractionated spinal cord samples were analyzed using a nano flow LC (Nano LC-2D, Eksigent Technologies) system coupled to a linear ion trap mass spectrometer (LTQ, ThermoFinnigan) following a 10 uL injection. The analytical column (Self-pack picofrit column, 75 μm ID, New Objective) was packed 15 cm with 3 μm C18 (Magic C18 AQ 200A 3U, Michrom Bioresources Inc). The trap column was obtained pre-packed from New Objective Inc. (Integrafrit sample trap, C18 5 μm, 100 μm column ID). The samples were trapped at an isocratic flow rate of 2 μl/min for 10 minutes and eluted at a flow rate of 300 nl/min via a mobile phase gradient of 2 – 50% B in 180 min (Mobile Phase A: 0.1% formic acid in water, Mobile Phase B: 0.1% formic acid in acetonitrile). The peptides were detected in the positive mode and the mass range for data acquisition was set from m/z 400-2000. The data were collected in Top 6 MS2 mode (N = 4) with the dynamic exclusion set for 30s, the exclusion size list set to 200, and the normalized collision energy for CID set to 35%. The capillary spray voltage was set to 2.5 kV. Peptide identification was performed with the SEQUEST algorithm with differential modification of methionine to its sulfoxide. The uniprotmus_frc.fasta mouse database, concatenated to a reversed decoy database, served to estimate a false discovery rate (FDR). Peptides were accepted within 1 Da of the expected mass, meeting a series of custom filters on ScoreFinal (S_f), -10 Log P, and charge state that attained an average peptide FDR of <2% across data sets. Manual inspection of spectra, FDR calculation,

and protein inference were performed in Proteomics Browser Suite 2.23 (ThermoFisher Scientific). We utilized an algorithm written in-house that reveals related MSMS spectra (MuQuest; Harvard Proteomics Browser Suite). To analyze the spinal cord peptidome to search for SP fragments, we used the MSMS data from the in vitro membrane lysate experiments. MuQuest is then applied to compare the in vitro MSMS spectra with those of the in vivo data set to determine which in vitro MSMS spectra (i.e., which peptides) are present in the in vivo samples. The output files are filtered based on charge state, mass to charge values, and statistical scores.

Isotope dilution MS (IDMS) to determine endogenous levels of substance P and its c-terminal peptide fragments

The heavy-label versions of SP(1-7), SP(1-9), and substance P were spiked into spinal cord samples at the beginning of the peptide isolation process. After fine tuning the amount of spiked peptides, it was determined that a final concentration of 100 fmol/μL of SP(1-7) and substance P and 25 fmol/μL of SP(1-9) would lead to adequate measurements of the peptides using a comparison of the area integration for the +2 charge state of the endogenous and heavy labeled peptides..

Monitoring C-terminal SP degradation in spinal cord lysates

Three mouse spinal cords were dounce homogenized in 1.1 mL phosphate buffered saline (PBS) on ice and then sonicated for 15 s at 4 ºC. Tissue debris was separated by centrifuging the sample at 5,000 x g for 20 min at 4 ºC. The soluble fraction was collected after ultracentrifuging the sample at 55,000 x g for 1 h at 4ºC. The membrane pellet that remained was washed 3x with 600 μL PBS by gently covering the pellet. The pellet was then suspended in 400 μL PBS. The protein content in the soluble and membrane lysate was quantified by the Bradford assay. Substance P (100 μM) was incubated in 1 mg/mL soluble and membrane lysates for 1 h at 37ºC. The reactions proceeded at 37ºC for 1 h and were quenched with an equal volume of 8 M guanidinium hydrochloride. The samples were then desalted via C18 ZipTips (Millipore) and speed vac dried. The samples were redissolved in 0.1% formic acid (aq) and analyzed by MALDI-TOF MS for substance P-degrading activity (i.e. formation of SP(1-7) and SP(1-9)) using the method outlined in “MALDI-TOF MS and LC-MS/MS analysis of *in vitro* peptides” section.

Developing a candidate list for the substance P degrading enzymes responsible for the formation of SP(1-9)

The MEROPS database was utilized to devise a candidate list for the enzymes that could cleave substance P, forming SP(1-9) in mice (32). The candidate list was narrowed based on protein abundance using the following databases:

www.brain-map.org

https://www.nextbio.com/b/nextbio.nb

Western blotting to confirm presence of candidate enzymes

The presence of the candidate enzymes in the mouse spinal cord membrane lysate was confirmed by Western blot using the following antibodies: matrix metallopreoteinase 8 (MMP8) (Abcam Inc.; rabbit polyclonal), matrix metallopreoteinase (MMP9) (Abcam Inc.; rabbit polyclonal), endothelin converting enzyme 1 (ECE1) (Abcam Inc.; rabbit polyclonal), endothelin converting enzyme 2 (ECE2), (Proteintech Group Inc.; rabbit polyclonal), pitrilysin (PITRM1) (Proteintech Group; rabbit polyclonal), neprilysin (CD10) (Abcam Inc.; mouse polyclonal). To improve sensitivity/detection, the membrane lysate was fractionated using the membrane protein solubilization properties of deoxycholate (Alfa Aesar) at different concentrations (1, 4, 12, and 24 mM).

Protease inhibitor studies using general enzyme class inhibitors

The membrane fraction of spinal cord lysates (1 mg/mL) were pre-incubated at 37 ºC for 30 minutes separately with each the following inhibitors: 10 μM E-64 (cysteine protease), 1 mM iodoacetamide (cysteine protease), 1 mM o-phenanthroline (metalloprotease), 10 μM pepstatin A (aspartyl protease), 1 mM phenylmethylsulfonyl fluoride (serine protease), 1 mM diisopropylfluorophosphate (serine protease), and vehicle (PBS with DMSO for 5% DMSO final concentration in reaction). After the pre-incubation with each inhibitor, substance P was added to 100 μM final concentration. The reactions proceeded at 37ºC for 1 h and were quenched with an equal volume of 8 M guanidinium hydrochloride. The samples were then desalted via C18 ZipTips (Millipore, Billerica, MA) and speed vac dried. The samples were redissolved in 0.1% formic acid (aq) and analyzed by LC-MS/MS for subtance P-degrading activity (i.e. formation of SP(1-7) and SP(1-9)) using the method outlined in "MALDI-TOF MS and LC-MS/MS analysis of *in vitro* peptides" section.

Protease inhibitor studies using metalloprotease specific enzyme class inhibitors

The membrane fraction of spinal cord lysates (1 mg/mL) were pre-incubated at 37 ºC for 30 minutes separately with each the following inhibitors at: SM-19712 (ECE1), phosphoramidon (neprilysin; ECE2), MMP9 inhibitor, TIMP2 (MMP

inhibitor), GM6001 (MMP and neprilysin broad inhibitor), chymostatin (pitrilysin), captopril (ACE), enalaprilat (ACE), actinonin (Meprin 1A), and vehicle (PBS with DMSO for 5% DMSO final concentration in reaction). All inhibitors were present at 100 μM except for TIMP2 which was present at 4 μM. After the pre-incubation with each inhibitor, either SP(1-11) or SP(1-9) was added to 100 μM final concentration. The reactions proceeded at 37ºC for 1 h and were quenched with an equal volume of 8 M guanidinium hydrochloride. The samples were then desalted via ZipTip C18 (Millipore, ZTC18S096) and speed vac dried. The samples were redissolved in 0.1% formic acid (aq) and analyzed by LC-MS/MS for SP(1-11) or SP(1-9)-degrading activity.

In vivo and in vitro comparative study of substance P degradation in WT and *Nep*$^{-/-}$ mice spinal cord tissue

The levels of substance P in WT and *Nep*$^{-/-}$ mice spinal cord tissue were compared by IDMS using Top 6 MS/MS as described previously. The degradation of substance P was compared in WT and *Nep*$^{-/-}$ mice spinal cord membrane lysate using LC-MS/MS analysis.

MALDI-TOF MS and LC-MS/MS analysis of *in vitro* peptides

MALDI-TOF MS was performed with α-cyano-4-hydroxycinnamic acid as the matrix using 2 μL of a 50 μM reconstituted degradation reaction solution (based on initial substance P concentration).

For LC-MS analysis, an Agilent 6220 LC-ESI-TOF instrument was used in the positive mode. A Bio-Bond C18 (5 μm, 150 x 2.1 mm) column was used together with a precolumn (C18, 3.5 μm, 2 x 20 mm). Following injection of 25 μL of 5 μM solutions, the samples were trapped at an isocratic flow rate of 0.1 ml/min for five minutes and eluted at a flow rate of 0.25 mL/min via a mobile phase gradient of 2 – 100% B in 40 min (Mobile Phase A: 0.1% formic acid in water, Mobile Phase B: 0.1% formic acid in acetonitrile). MS analysis was performed with an electrospray ionization (ESI) source. The capillary voltage was set at 4.0 kV and the fragmentor voltage to 100 V. The drying gas temperature was 350 °C, the drying gas flow rate was 10 L/min, and the nebulizer pressure was 45 psi. Data was collected in the centroid mode using a mass range of 100-1500 Da. The peptides were analyzed by mass extraction of the +3 charge state.

Using a rhodamine-tagged hydroxamate benzophenone probe (HxBP-Rh) to determine GM6001-binding proteins in mice membrane lysates.

Standard conditions for HxBP-labeling reactions were as follows. Extracted membrane lysates (6 mg/mL) were mixed with 100 nM of HxBP-Rh in the presence or absence of 100 μM GM6001. The mixtures were preincubated on ice for 30 min before irradiation at 365 nm for 1 h (on ice) followed by quenching with 1/2 vol of standard 3X SDS/PAGE loading buffer (reducing). A control was performed in the absence of GM6001 and no irradiation. Labeled proteins were

visualized in-gel with fluorescence scanning using a Typhoon flatbed fluorescence scanner (GE Healthcare Life Sciences).

GM6001 injection experiments

For GM6001 injection experiments, 3-4 month old female WT mice (n=4) were fasted overnight. GM6001 was dissolved at a high concentration in DMSO. Injections were performed intraperitoneally with a 10 µL/g injection of either vehicle (5% DMSO, 95% saline) or 10 mg/mL GM6001 in 5% DMSO, 95% saline for a final dose of 100 mg/kg GM6001. Mice were allowed to return to their cages for three hours and then spinal cords were isolated as described in the 'Animal studies' section.

3.12 References

1. Otsuka M & Yoshioka K (1993) Neurotransmitter functions of mammalian tachykinins. *Physiol Rev* 73(2):229-308.

2. Hall ME & Stewart JM (1983) Substance P and antinociception. *Peptides* 4(1):31-35.

3. Hall ME & Stewart JM (1983) Substance P and behavior: opposite effects of N-terminal and C-terminal fragments. *Peptides* 4(5):763-768.

4. Yeomans DC & Proudfit HK (1992) Antinociception induced by microinjection of substance P into the A7 catecholamine cell group in the rat. *Neuroscience* 49(3):681-691.

5. Kolasinski SL, Haines KA, Siegel EL, Cronstein BN, & Abramson SB (1992) Neuropeptides and inflammation. A somatostatin analog as a selective antagonist of neutrophil activation by substance P. *Arthritis Rheum* 35(4):369-375.

6. Palmer JM & Greenwood B (1993) Regional content of enteric substance P and vasoactive intestinal peptide during intestinal inflammation in the parasitized ferret. *Neuropeptides* 25(2):95-103.

7. Berrettini WH*, et al.* (1985) CSF substance P immunoreactivity in affective disorders. *Biol Psychiatry* 20(9):965-970.

8. Shirayama Y, Mitsushio H, Takashima M, Ichikawa H, & Takahashi K (1996) Reduction of substance P after chronic antidepressants treatment in the striatum, substantia nigra and amygdala of the rat. *Brain Res* 739(1-2):70-78.

9. Lieb K*, et al.* (2002) Effects of the neuropeptide substance P on sleep, mood, and neuroendocrine measures in healthy young men. *Neuropsychopharmacology* 27(6):1041-1049.

10. Blumberg S, Teichberg VI, Charli JL, Hersh LB, & McKelvy JF (1980) Cleavage of substance P to an N-terminal tetrapeptide and a C-terminal heptapeptide by a post-proline Cleaving enzyme from bovine brain. *Brain Res* 192(2):477-486.

11. Nolte WM, Tagore DM, Lane WS, & Saghatelian A (2009) Peptidomics of prolyl endopeptidase in the central nervous system. *Biochemistry* 48(50):11971-11981.

12. Tenorio-Laranga J, Valero ML, Mannisto PT, Sanchez del Pino M, & Garcia-Horsman JA (2009) Combination of snap freezing, differential pH two-dimensional reverse-phase high-performance liquid chromatography, and iTRAQ technology for the peptidomic analysis of the effect of prolyl oligopeptidase inhibition in the rat brain. *Anal Biochem* 393(1):80-87.

13. Tinoco AD*, et al.* (2011) A peptidomics strategy to elucidate the proteolytic pathways that inactivate peptide hormones. *Biochemistry* 50(12):2213-2222.

14. Tagore DM*, et al.* (2009) Peptidase substrates via global peptide profiling. *Nat Chem Biol* 5(1):23-25.

15. Tinoco AD, Tagore DM, & Saghatelian A (2010) Expanding the dipeptidyl peptidase 4-regulated peptidome via an optimized peptidomics platform. *J Am Chem Soc* 132(11):3819-3830.

16. Andren PE & Caprioli RM (1995) In vivo metabolism of substance P in rat striatum utilizing microdialysis liquid chromatography microelectrospray mass spectrometry. *Journal of Mass Spectrometry* 30(6):817-824.

17. Lee CM, Campbell NJ, Williams BJ, & Iversen LL (1986) Multiple tachykinin binding sites in peripheral tissues and in brain. *Eur J Pharmacol* 130(3):209-217.

18. Huang RR, Yu H, Strader CD, & Fong TM (1994) Interaction of substance P with the second and seventh transmembrane domains of the neurokinin-1 receptor. *Biochemistry* 33(10):3007-3013.

19. Rawlings ND, Barrett AJ, & Bateman A (2012) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res* 40(Database issue):D343-350.

20. Defendini R & Zimmerman EA (1993) Angiotensin converting enzyme in the brain. *J Neurochem* 60(2):787-789.

21. Yokosawa H, Endo S, Ogura Y, & Ishii S (1983) A new feature of angiotensin-converting enzyme in the brain: hydrolysis of substance P. *Biochem Biophys Res Commun* 116(2):735-742.

22. Defendini R, Zimmerman EA, Weare JA, Alhenc-Gelas F, & Erdos EG (1983) Angiotensin-converting enzyme in epithelial and neuroepithelial cells. *Neuroendocrinology* 37(1):32-40.

23. Defendini R, Zimmerman EA, Weare JA, Alhenc-Gelas F, & Edros EG (1982) Hydrolysis of enkephalins by human converting enzyme and localization of the enzyme in neuronal components of the brain. *Adv Biochem Psychopharmacol* 33:271-280.

24. Matsas R, Rattray M, Kenny AJ, & Turner AJ (1985) The metabolism of neuropeptides. Endopeptidase-24.11 in human synaptic membrane preparations hydrolyses substance P. *Biochem J* 228(2):487-492.

25. Kato T*, et al.* (1978) Successive cleavage of N-terminal Arg1--Pro2 and Lys3-Pro4 from substance P but no release of Arg1-Pro2 from bradykinin, by X-Pro dipeptidyl-aminopeptidase. *Biochim Biophys Acta* 525(2):417-422.

26. Horsthemke B*, et al.* (1984) Subcellular distribution of particle-bound neutral peptidases capable of hydrolyzing gonadoliberin, thyroliberin, enkephalin and substance P. *Eur J Biochem* 139(2):315-320.

27. Lee CM, Sandberg BE, Hanley MR, & Iversen LL (1981) Purification and characterisation of a membrane-bound substance-P-degrading enzyme from human brain. *Eur J Biochem* 114(2):315-327.

28. Probert L & Hanley MR (1987) The immunocytochemical localisation of 'substance-P-degrading enzyme' within the rat spinal cord. *Neurosci Lett* 78(2):132-137.

29. Lein ES*, et al.* (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445(7124):168-176.

30. Saghatelian A, Jessani N, Joseph A, Humphrey M, & Cravatt BF (2004) Activity-based probes for the proteomic profiling of metalloproteases. *Proc. Natl. Acad. Sci. U. S. A.* 101(27):10000-10005.

31. Tinoco AD*, et al.* (2011) A Peptidomics Strategy To Elucidate the Proteolytic Pathways That Inactivate Peptide Hormones. *Biochemistry* 50(12):2213-2222.

32. Rawlings ND, Barrett AJ, & Bateman A (2012) MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research* 40(D1):D343-350.